1. Introduction

Part-of-speech tagging (POS tagging), also called grammatical tagging or word-category disambiguation, is the process of categorizing (marking up) a word in a text (corpus) as corresponding to a certain part of speech, based on both its ontology and its context.

Hidden Markov Models, as a particular instance of directed graphical model (also called Bayesian Network), are widely used for predicting sequential data. Also, people also take advantage of Conditional Random Fields as undirected graphical model (also named Markov Random Field) for the task of sequential data prediction.

In this experiment, we will examine the predictive powers of Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) on the POS tagging task for the further study of predecessors' work (1) and (2).

2. Experiment

The experimented datasets are *atis* and *wsj*, both of which are sourced from Linguistic Data Consortium (LDC)'s Penn Treebank collection. Since the entire *wsj* dataset is too large, we will train the model by its section 00 and test our model by its section 01. By default, every dataset will be split into training set and testing set with a ratio of $8:2$.

As to the metrics to assess the performance of our approaches, we measure the prediction accuracy as the percentage of items whose labels are correctly predicted. To make the performance evaluation less vulnerable to random train/test spliting, we will also measure the accuracy specifically for out-of-vocabulary items as a more fair performance indicator.

In the task of POS tagging, it turns out that adding extra orthographic features does help to improve the accuracy. In our experiments, orthographic features like capitalization (caps) as well as common English suffixes will be employed to assist the HMM-based and CRF-based POS tagger.

Table 1: Comparisons of Out-of-vocabulary Accuracies and Runtimes Between POS Models. All CRF models are trained with 500 iterations. We average 10 random split for statistics of atis3.

Models Datasets	atis				wsj			
POS Models	training	testing	OOV	runtime	training	testing	OOV	runtime
HMM	88.86%	86.62%	21.80%	8.366	82.96%	76.99%	36.25%	54.71
CRF-simple	99.88%	92.62%	25.75%	87.30	99.52%	80.40%	47.57%	10376
CRF-hyphen	99.88%	92.62%	25.75%	89.19	99.53%	80.72%	49.36%	11742
CRF-caps	99.88%	92.82%	29.42%	86.21	99.53%	80.40%	48.23%	11847
CRF-suffix	99.88%	93.28%	34.38%	89.42	95.25%	79.94%	50.05%	10699
CRF-all	99.88%	93.40%	38.61%	84.17	99.51%	83.67%	60.92%	10854

We implement several POS models for comparisons. HMM is the Hidden Markov Model solver built in the Mallet Library. CRF-simple is the Conditional Random Field Model with the no orthographic features fed. CRF-caps includes only the capitalization feature, CRF-suffix only includes "-s" and "-ing" features, and CRF-hyphen is fed with features recognizing hyphens and numbers. On top of that, CRF-all denotes the CRF model trained with all othographic features stated above.

3. Discussions

(a) How does the overall test accuracy of CRF and HMM differ (when using only tokens) and why?

It turns out that all CRF models (even CRF-simple, the one with no orthographic features) have higher testing accuracies than HMM model in either atis3 and selected wsj datasets. This could imply that CRF model basically has stronger predictive power for POS tagging task. The possible reason for this is that CRF, as a generative model, could capture a more complex structural relations between words and tags (for example, CRF handles calibrated probability estimates whereas HMM only use occurrence counts for training).

(b) How does the test accuracy for OOV items for CRF and HMM differ (when using only tokens) and why?

It can be observed that all CRF models (even CRF-simple, the one with no orthographic features) have higher testing OOV accuracies. The possible reason is that CRF has better predictive power that can be generalized to unseen instances in training stages.

(c) How does the training accuracy of HMM and CRF differ and why? The training accuracies of CRF models are all higher than those of HMM models in both experimental datasets. This might be attributed to the dintinguished nature of generative model and discriminative model. HMM, as discrminative model, could only capture linear relations between words and tags, where as CRF can capture more complex relations.

(d) How does the run time of HMM and CRF differ and why?

The run time of training CRF is much much longer than that of training HMM models. This can be explained as follows. The CRF model handles calibrated probability estimates and use certain convex optimization to iteratively find optimal parameters. However, HMM only use occurrence counts and their division to figure out its model parameters (not iterative).

(e) How does adding orthographic features affect the accuracy (both overall and OOV) and runtime of the CRF and why?

It is obvious that adding orthographic features generally increase testing accuracy and especially OOV accuracy. Another noticeable observationis is that those orthographic features for CRF models do not increase or decrease the training time for CRF models. This is because CRF model is undirected graphical model and its training algorithm can easily tolerate high-dimensional input.

(f) Which features helped the most? (i.e. try only including some feature types and not others)

Comparing CRF-hyphen, CRF-caps, CRF-suffix with CRF-simple on OOV accuracy column, it can be found that suffix information (-s and -ing) help most for both the atis3 dataset and the selected wsj dataset. For wsj dataset, we can see that CRF-suffix well cope with overfitting issue and gives a good generalization performance (OOV accuracy).

References

[1] Lafferty, J., McCallum, A., and Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[2] Sutton, C., and McCallum, A. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning* (2006), 93–128.