Answers - CS 545, Fall 2012, Homework 4

Kumaresh Visakan Murugan
mvisakan@cs.wisc.edu

1. Answer to Question 1

The simple baseline has been implemented and the overall accuracy and the unseen accuracy are calculated using this model. The accuracies are as follows:

Language	Overall Accuracy	Unseen Word Accuracy
English	93.1320487534%	48.6931268151%
Czech	85.4976357975%	0.0%
Polish	89.9469247580%	39.7177419355%

The strange thing observed is that the unseen word accuracy for Czech language is 0%. This happens because the most common tag overall found in Czech is PUN and all the unseen words are mapped incorrectly to this default tag.

2. Answer to Question 2

The Bigram HMM model has been implemented and the overall accuracy and the unseen word accuracy are now found to be:

Language	Overall Accuracy	Unseen Word Accuracy
English	94.4773175542%	42.6911907067%
Czech	88.5192019375%	21.8804159445%
Polish	87.1120824227%	21.4516129032%

It is found that on comparison with the simple baseline implemented before,

- The overall accuracy for English has increased by a small value, but the unseen word accuracy has decreased.
    - The Bigram HMM model results in correctly identifying the tags of seen words and thus increases the overall accuracy, but has not correctly identified the tags for the unseen words when compared with the baseline implementation.
- The overall accuracy and unseen word accuracy for Czech has both increased
    - The new model results in identifying correct tags for some unseen words, as the unseen words are no longer tagged using the default tag PUN. This has resulted in the increase of the overall accuracy of the model as well.
- The overall accuracy and unseen word accuracy for Polish has both decreased
    - The Bigram HMM model is based on the Markov assumption that the tags depend on the previous tag and also on the word. For polish it seems that this first order Markov assumption was not able to fit the data correctly. Hence the accuracies have decreased.

3. Answer to Question 3

The HMM model used before is changed to bias the model so that it is more likely to use the overall most frequent tag when a word is unobserved in the training. The overall accuracy and the unseen word accuracies are now found to be:

Language	Overall Accuracy	Unseen Word Accuracy
English	94.7352450311%	48.6931268151%
Czech	85.7052243109%	0.9965337955%
Polish	89.8157976897%	38.9112903226%

- The overall accuracy and the unseen word accuracy for the 3 languages using the biased HM model are very much similar to the baseline model implemented before.

- This change in smoothing seems to have biased the model to predict the most frequent tag for the unseen words, hence deriving accuracies similar to the simple baseline model implemented before.

- Bayesian Interpretation of this estimation procedure :

  The Bayesian HMM model places prior multinomial distribution over the emission probabilities. This distribution forms the parameter of the model. The tag sequence for the words are predicted using the transition probability and the emission probability. We use symmetric Dirichlet priors for the distribution of emission probabilities.

  $p(t|w, \Theta) = (count(t, t') / count(t)) * ((count(t, w) + \Theta) / (count(t) + |V| * \Theta))$
  where Θ is the parameter of the prior over the emission distribution, t is the tag to be predicted for word w and t' is the previous tag.

  The transition and emission probabilities are the priors used to calculate the posterior probability of the tag corresponding to a particular word based on the previous tag.

  Different concentration parameters are used for different emission distributions. For the most frequent tags a higher concentration parameter is used as a prior whereas a lower concentration parameter is used for other tags over unseen words.
  In case of frequent tags, the parameter is,

  $$\Theta = 1000 * \delta$$

  In case of rest of the tags, the parameter is,

  $$\Theta = \delta$$

4. Answer to Question 4

The linear sequence model has been implemented and the overall accuracy and the unseen word accuracy are found to be:

Language	Overall Accuracy	Unseen Word Accuracy
English	92.6920548222%	53.1461761859%
Czech	87.4754930227%	40.7712305026%
Polish	90.3278176709%	48.3467741935%

- **English**
  - The overall accuracy has decreased while the unseen word accuracy has increased when compared with the previous HMM model.
- **Czech**
  - The overall accuracy has increased by a small value but the unseen word accuracy has significantly increased when compared with the previous HMM model.
- **Polish**
  - Both the overall accuracy and the unseen word accuracy has increased when compared with the previous HMM model

5. Answer to Question 5

Now let us implement the average perceptron in the weights learning phase of the linear sequence model implemented before.

The averaged vector constructed can be expressed as,

$$\lambda = (1 / t) * [\lambda^{(1)} + \ldots + \lambda^{(t)}]$$

where t = number of iterations X number of training examples. We can discard T as it will not have any effect on argmax predictions

- Expressing each weight vector $\lambda^{(t)}$ in terms of the previous weight vector and the perceptron update $\Delta^{(t)}$,

$$\lambda^{(1)} = \lambda^{(0)} + \Delta^{(1)}$$

$$\lambda^{(2)} = \lambda^{(1)} + \Delta^{(2)}$$

.

.

.

$$\lambda^{(t)} = \lambda^{(t-1)} + \Delta^{(t)}$$

- Now unrolling the recursion and writing each weight vector in terms of the updates,

$$\lambda^{(1)} = \Delta^{(1)}, \text{ as } \lambda^{(0)} = 0$$

$$\lambda^{(2)} = \Delta^{(1)} + \Delta^{(2)}$$

.
.
.

$$\lambda^{(t)} = \Delta^{(1)} + \ldots + \Delta^{(t)}$$

- Plugging these values to the definition of λ,

$$\lambda = [t\,\Delta^{(1)} + (t-1)\,\Delta^{(2)} + \ldots + \Delta^{(t)}]$$

- The overall accuracy and the unseen accuracy are now found to be:

Language	Overall Accuracy	Unseen Word Accuracy
English	95.1347797502%	66.5053242982%
Czech	91.3504786069%	52.7729636049%
Polish	90.7024664377%	51.8145161290%

- The overall accuracy and the unseen word accuracy for the 3 languages have increased when compared with the previous linear sequence model implemented using regular perceptron.

6. Answer to Question 6

The linear sequence model is improved by adding additional features by checking at the suffix of the words. The overall and unseen accuracy are now found to be:

Language	Overall Accuracy	Unseen Word Accuracy
English	96.1310878471%	81.5101645692%
Czech	96.4248644908%	85.7019064125%
Polish	97.3587261942%	90.2822580645%

- The overall accuracy and the unseen word accuracy for the 3 languages have increased significantly when compared with the previous results.
- The linear sequence model with additional features checking the suffices of the words implemented using averaged perceptron achieves the highest performance and the accuracies calculated so far indicates the same.