NLP Homework-3

Part-1

Question 1.1 :

The advantage of using a smaller tagset is lower number of training instances would be required for the model to be trained , therefore it will in turn take lower time for convergence.

The disadvantage of using a smaller tagset is that the accuracy of the model on unseen data would be reduced as we would not be able to generalize very well for this data considering that we have only seen 17 of the 45 tags in our training data.

Part-2

Question 2.1 : The training data does not contain all possible trigrams that can appear in the test data. Therefore when a trigram is not seen in the training data we can backoff to the skip bigram or in our case look if the skip bigram matches with the training data and thus make our predictions on unseen trigrams more accurate. The advantage of using skip-bigram over trigram because sometimes the word immediately before a word conveys little information about the tag of the current word. For example, consider the sentence " She wore a beautiful dress." In order to predict the tag of the word beautiful, the trigram "wore a" conveys equivalent if not less information when compared to the skip-bigram "wore" (considering the article "a" gives no information about the tag of beautiful).

Question 2.2 : The last set of features, in this case referring to the word features are extremely useful. These features include capital, hyphen, digit , suffixes and prefixes give important information about part of speech the given word belongs to. For example, generally if a word starts with a capital (if it is not the starting letter of a sentence) then it is most likely a noun. Further, if a word ends with a suffix "ed" , it is most likely a verb.

Part-3

Question 3.1 : The reasons for removing rare features are :

1. The rare features occur for less than five words, for accommodating these features the size of the matrix would be increased, which in turn would lead to an increase in the time required for training the given model.
2. Including rare features in the training data, would lead to overfitting. The model would be very specific to the training data and would not be able to generalize. Therefore, when the model tries to predict for unseen test data, it' s accuracy would be reduced. This is the reason we remove rare features which are so rare that they seem very specific to the given training instances.

Question 3.2: We want to use sparse matrix for X train because there are many features in the training data (many columns in the matrix) but each training instance (in our case, a word) has a very few of those features which means that the values of all the other columns other than those few columns would be zeros. To avoid storing a large matrix, mostly filled with zeros we chose to use a sparse matrix that only stores the column index and the row index of the non-zero values of the matrix. The advantage of a sparse matrix over a dense matrix is that it occupies less memory in storage which in turn reduces the computational requirement of the system.

Part-4

Question 4.2 : The predicted tag sequences for each of the test sentences is as follows :

1. ['ADJ', 'NOUN', 'VERB', 'DET', 'ADJ', 'ADJ', 'NOUN', 'NOUN', 'VERB', 'ADP', 'NOUN', '.', 'NOUN', '.']
2. ['DET', 'ADJ', 'NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'ADJ', 'NOUN', '.']
3. ['NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'ADJ', 'NOUN', 'ADP', 'DET', 'NOUN', 'PRT', 'VERB', 'DET', 'ADJ', 'NOUN', 'ADP', 'NOUN','.']
4. ['PRON', 'VERB', 'PRT', 'VERB', 'ADP', 'DET', 'NOUN', '.']
5. ['NOUN', 'VERB', 'VERB', 'ADP', 'NUM', '.']

Part-5

Question 5.1 : This homework took me about 30 hours to finish.

Question 5.2 : I discussed this homework with Nikhil.

Part-6

Question 6.1 : The new predicted tag sequence for each of the test sentences is as follows :

1. ['NOUN', 'NOUN', 'VERB', 'DET', 'ADJ', 'ADJ', 'NOUN', 'NOUN', 'VERB', 'ADP', 'NOUN', '.', 'NOUN','.']
2. ['DET', 'ADJ', 'NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'ADJ', 'NOUN', '.']
3. ['NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'NOUN', 'NOUN', 'ADP', 'DET', 'ADJ', 'PRT', 'VERB', 'DET', 'NOUN', 'NOUN', 'ADP', 'NOUN', '.']
4. ['PRON', 'VERB', 'PRT', 'VERB', 'ADP', 'DET', 'NOUN', '.']
5. ['NOUN', 'VERB', 'VERB', 'ADP', 'NUM', '.']

Most of the tags have remained the same when compared to the previous predicted tag sequences. However, they are a few changes. For example, the first tag in the first sentence is NOUN which is the correct tag for Apple in the phrase 'Apple Inc.'. Further the eighth tag in the 3rd sentence changed to NOUN which is the correct tag for the word "Trump". Thus, there have been a few changes that show that the predictions are slightly better than the predicted tags in part 4.