Task 1:

What threshold value did you choose for identifying unknown words for replacement?

There are 43193 unique words in the training data. 20011 words have an occurrence frequency of less than 2, nearly half of the unique words. Thus, I chose 2 as my threshold value for identifying unknown words for replacement.

What is the overall size of your vocabulary, and how many times does the special token ”< unk >” occur following the replacement process?

The overall size of my vocabulary is 23183. The special token “<unk>” occurs 20011 times.

Task 2:

How many transition and emission parameters in your HMM?

There are 2025 transition parameters and 1043235 emission parameters.

Task 3:

What is the accuracy on the dev data?

The accuracy on the dev data is about 93.4%.

Task 3:

What is the accuracy on the dev data?

The accuracy on the dev data is about 94.7%.