HMM POS Tagging Heuristics

My HMM probabilities and Viterbi's prob and prev variables are stored in dictionary data structures. For the recursive case of the Viterbi algorithm, the argmax_x line of the Viterbi pseudocode is implemented in a way that it only goes through the previous timestep (word) that has probability value of larger than 0 in prob[t-1]. Probabilities that are 0 are also not stored in the prob or prev variables since they do not provide any use to be stored. These are done so to help for a faster running time.

On top of the Viterbi algorithm, I implemented a probability distribution for test words that are not present in the training data. This probability distribution uses the previous two most likely tags to find such combination of tags in the training data, records the preceding tag and its occurrence, and create a probability for each possible preceding tag. This will then act as the prob probability of the Viterbi algorithm. The prev value to this tag will be the tag with (t-1)'s largest prob value. If this heuristic does not produce any probabilities, another probability distribution will be created, but instead of using the previous two tags, it only looks the one previous tag before the unknown word. Keeping previous 2 as the first approach has resulted from comparing the accuracy results of different runs. As written in the table below, the accuracy increases as lower number of previous tags are used. I did not end up using only 1 previous tag since the trade-off between accuracy and runtime is not worth it. One previous tag run may run up to 12 minutes.

For first sentence words that are not present in the training data, the prob value will equal to the I probability.

Next, I notice a lot of names are tagged incorrectly. So, I implemented a heuristic where if the word has been tagged as NP0 in previous sentences, it starts with an uppercase letter, and all the characters are alphabet letters, the word should be tagged as NP0 too.

Test data	Using Training data	Limit to at most 3 prev tags	Limit to at most 2 prev tags	Limit to at most 1 prev tags	Limit to at most 2 prev tags and NP0
1	1				96.34%
1	3		83.74%	83.88%	83.94%
1	4		81.30%		83.13%
1	5	84.09%	84.54%	84.61%	86.09%
3	2				84.44%
5	1		84.18%		