Part-of-Speech Tagging Output

Here, I built a part-of-speech Hidden Markov Model (HMM), training on the first 10,000 sentences of the Brown Corpus. For the first part of the question, in the file *generate_components_hmm.py*, I build the three main parts of the HMM: the transition matrix, the initial state probabilities, and the observation matrix. I also hung onto the mappings between state/observations and indices Subsequently, I moved into the next step to implement the model by creating Viterbi and backpointer matrices in the *viterbi.py* file. In this file I tested the function on three test sentences from the Brown Corpus (never seen by the model). The three sentences were:

- Those coming from other denominations will welcome the opportunity to become informed.
- The preparatory class is an introductory face-to-face group in which new members become aquainted with one another.
- It provides a natural transition into the life of the local church and its organizations.

Out of the 47 words in these tests, I was able to correctly tag 44 (93.6% accuracy). First, the model miss-tagged the word 'coming' in the first setence. "Coming" was tagged as noun when it should be tagged as a verb. Next, the program missed two words in the second sentence. "Face-to-face" was an unknown word for the model—it was tagged as a noun, but is actually an adjective in context. Finally, "another" was tagged as a noun when it should have been tagged as a determiner. In the third sentence, the sequence was tagged entirely correct.

In order to try to identify why the program failed in these cases, I dug a little deeper into these missed words. First, for the word "coming", I noticed that it follows a determiner. From the transition matrix, I can see that the probability of a noun given a determiner is 0.62, while the probability of a verb given a determiner is only 0.06. The model likely performed incorrectly here because this is an unlikely sentence organization. For "face-to-face", I can see that it follows an adjective, and because it is an unknown word, the model does not have anything to calcualte probabilities on besides the previous state, and nouns frequently follow adjectives (.67). Finally, for the last missed word, "another" follows a number, "one." A determiner following a number is also a very unlikely situation, with a probability of 0.01. All in all, it seems that the model generally performs very well at tagging POS, and generally makes mistakes in two cases: (1) when there are unknown words and (2) when there are uncommon sentence structures.