Computational Linguistics, Summer 2015, Exercise “POS-Tagging”

1. Run the viterbi algorithm (using pen and paper) for the input “man the boat” and the following (partically specified) HMM:

Emission probabilities:

	DT	**NN**	**VB**	**...**
man	0.05	0.7	0.1	...
the	0.4	0.05	0.05	...
boat	0.05	0.8	0.05	...
...	...	...	...	...

Transition probabilities:

	DT	**NN**	**VB**	**End**	**...**
DT	0.05	0.7	0.1	0.1	...
NN	0.2	0.15	0.3	0.2	...
VB	0.3	0.2	0.1	0.2	...
Start	0.4	0.1	0.2	0.0	...
...	...	...	...	...	...

2. Complete the implementation of the HMM-tagger which you can download from the course website. You have to implement the two following two methods

P_emit(self, y, o) - the emission probability of the word o given the tag y

P_trans(self, x, y) - the transition probability of the tag y given tag x

In a first step, implement a “strict” version of the probabilities, i.e., the two methods should return 0 in case a word or bigram has not been seen in the training data.

3. Evaluate your tagger implementation.

a. How many words receive their correct tags?

b. How many sentences receive a non-zero probability?

c. How many words receive their correct tags if you take only known words into account, i.e., if you evaluate only on sentences with a non-zero probability?

4. Implement One-Count smoothing to estimate probabilities of unknown words. The method is described in Jason Eisners tagging tutorial (see course website).

5. Redo the evaluation.