Hw6

Format of HMM

```
state_num=nn            ## the number of states
sym_num=nn              ## the size of output symbol alphabet
init_line_num=nn        ## the number of lines for the initial probability
trans_line_num=nn       ## the number of lines for the transition probability
emiss_line_num=nn       ## the number of lines for the emission probability

\init
state  prob    lg_prob                  ## prob=\pi(state), lg_prob=lg(prob)
  ...

\transition
from_state  to_state   prob   lg_prob      ## prob=P(to_state | from_state)
  ...

\emission
state  symbol  prob   lg_prob              ## prob=P(symbol | state)
  ...
```

Q1: HMM for a bigram tagger

- cat training_data | create_2gram_hmm.sh output_hmm
- training data: w1/t1 … wn/tn
- No smoothing

Q2: HMM for a trigram tagger

- cat training_data | create_3gram_hmm.sh output_hmm l1 l2 l3 unk_prob_file

- unk_prob_file: “tag prob”: P(<unk>|tag)=prob

- Smoothing:

$$P_{int}(w_3 \mid w_1, w_2) = \lambda_3 P_3(w_3 \mid w_1, w_2) + \lambda_2 P_2(w_3|w_2) + \lambda_1 P_1(w_3)$$

$$P_{smooth}(w \mid tag) = P(w \mid tag) * (1 - P(< unk >| tag))$$

Q3: check HMM

- check_hmm.sh input_hmm > warning_file

state_num=6
sym_num=11
warning: different numbers of init_line_num: claimed=2, real=1
warning: different numbers of trans_line_num: claimed=13, real=15
warning: different numbers of emission_line_num: claimed=11, real=12
warning: the trans_prob_sum for state N is 0.9
warning: the trans_prob_sum for state V is 1.1
warning: the emiss_prob_sum for state BOS is 0
warning: the emiss_prob_sum for state N is 0.5
warning: the emiss_prob_sum for state V is 0.85
warning: the emiss_prob_sum for state Adv is 0