Readme

Yao Yan
1669598

Q1. create_2gram_hmm.sh

- Add bos/BOS and eos/EOS to the beginning and the end of each sentence
- Separate the word and tag by '/', but there is an exception (Macmillan\/McGraw-Hill/NNP), if '\/' in the sentence, '/' is treated differently.
- Create some dictionaries for saving the information:
    - d_state: state; occurrence of the state like (DT, 30)
    - d_trans: ((from_state, to_state), count), like ((DT, Adj),20)
    - d_emiss: ((state, word), count), like ((DT, a),21)
    - d_symbol (word,count) like (we, 2)

Q2.create_3gram_hmm.sh

- Add bos/BOS and eos/EOS to the beginning and the end of each sentence
- Separate the word and tag by '/', but there is an exception (Macmillan\/McGraw-Hill/NNP), if '\/' in the sentence, '/' is treated differently.
- Create some dictionaries for saving the information:
    - d_state: state; occurrence of the state like (DT, 30)
    - d_bi: ((t1,t2), #) like ((DT NN),23)
    - d_tri: ((t1,t2,t3), #) like ((DT Adj NN),23)
    - p(t3 | t1, t2) = l3* $\frac{dtri(t1,t2,t3)}{dbi(t1,t2)}$ + l2* $\frac{dbi(t2,t3)}{dstate(t2)}$ + l1* $\frac{dstate(t3)}{all}$
    - d_trans: ((A_B, B_C), count), like ((DT_Adj, Adj_NN),20)
    - d_emiss: ((B_C, word), count), like ((DT_Adj, pretty),21)
    - d_symbol (word,count) like (we, 2)
    - p(w|tag) = p(w|tag) * (1- p(<unk>| tag))
    - There aren't unknown (t1,t2) because the probability of bigram are extracted from the same file where this output is generated

Q3.check_hmm.sh

- Check the claimed line number , symbol number and state number is correct or not
- Check the sum of the emission and transition whose first field are the same. If abs(sum - 1) > 0.001, print out warning.
- Print out warning for state in transition but not appearing in emission
- I used python dictionary to store states the HMM,
- d_trans: ((A_B), sum_of_probability), like ((DT_Adj,0.23) (A_B) is from state
- d_emiss: ((B_C),sum_of_probability), like (DT_Adj, 0.45)
- Python list for storing state and symbol
- l_symbol ['we', 'like' …]; l_state [DT,Adj...]