Name:_____

Question 1: Write out the joint probability for the following sentence using the chain rule:

p(There, is, only, one, person, who, is, not, ordinary)

Answer:

p(There, is, only, one, person, who, is, not, ordinary)
=p(There)p(is|There)p(only|There,is)p(one|There,is,only)p(person|There,is,only,one)p(who|There,is,only,one,person)p(is|There,is,only,one,person,who)p(not|There,is,only,one,person,who,is)p(ordinary|There,is,only,one,person,who,is,not)

Write out the probability above using the second-order Markov assumption.

Answer:

p(There, is, only, one, person, who, is, not, ordinary)
=p(There)p(is|There)p(only|There,is)p(one|is,only)p(person|only,one)p(who|one,person)p(is|person,who)p(not|who,is)p(ordinary|is,not)

Question 2: Consider the following training corpus **T** of sentences:

- *START Karlsson is round STOP*
- *START He lives on the roof STOP*
- *START He is happy STOP*
- *START On the roof STOP*
- *START Karlsson lives happily STOP*

(b) Compute the following maximum likelihood parameters:

p(Karlsson|START)=count(START,Karlsson)/count(START)=2/5

p(Karlsson|lives,happily)=count(lives,happily,Karlsson)/count(lives,happily)=0/1

p(STOP|happy)=count(happy,STOP)/count(happy)=1/1

(c) Compute the probability of the following sentences under the trigram model trained on **T**:

START Karlsson is happy STOP

Answer:

p(START,Karlsson,is,happy,STOP)
=p(Karlsson|START)p(is|START,Karlsson)p(happy|Karlsson,is)p(STOP|is,happy)
=count(START,Karlsson)/count(START)*count(START,Karlsson,is)/count(START,Karlsson)*count(Karlsson,is,happy)/count(Karlsson,is)*count(is,happy,STOP)/count(is,happy)
=2/5*1/2*0/1*1/1=0

START Karlsson lives on the roof STOP

Answer:

p(START,Karlsson,lives,on,the,roof,STOP)
=p(Karlsson|START)p(lives|START,Karlsson)p(on|Karlsson,lives)p(the|lives,on)p(roof|on,the)p(STOP|the,roof)
=count(START,Karlsson)/count(START)*count(START,Karlsson,lives)/count(START,Karlsson)*count(Karlsson,lives,on)/count(Karlsson,lives)*count(lives,on,the)/count(lives,on)*count(on,the,roof)/count(on,the)*count(the,roof,STOP)/count(the,roof)
=2/5*1/2*0/1*1/1*2/2*2/2=0

Question 3: We have the following training corpus:
The green book STOP
My blue book STOP
His green house STOP
Book STOP

Assume we have a language model based on this corpus using linear interpolation with $\lambda=1/3$ for all *i*. Compute the value of the parameter p(book|the green) under this model. Assume STOP as part of your unigram model.

Answer:

total_words=14
p(book|the green)
=1/3*count(the,green,book)/count(the,green)+1/3*count(green,book)/count(green)+1/3*count(book)/total_words=1/3*1+1/3*1/2+1/3*3/14=1/3+1/6+1/14