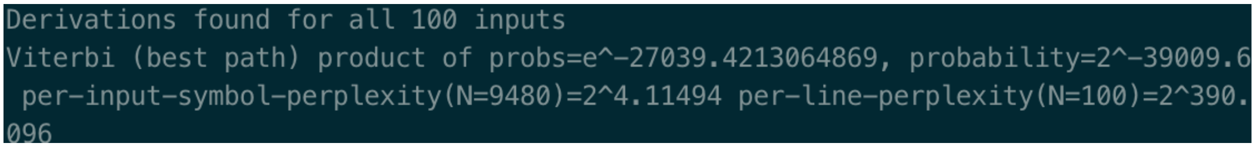
EX 2: Language Models and Entropy

Jing Huang (933039277)

1. Training Language Models

2.

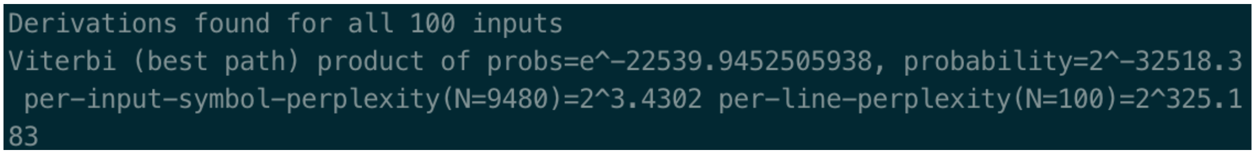
unigram:



```
Derivations found for all 100 inputs
Viterbi (best path) product of probs=e^-27039.4213064869, probability=2^-39009.6
 per-input-symbol-perplexity(N=9480)=2^4.11494 per-line-perplexity(N=100)=2^390.
096
```

entropy: 4.11494

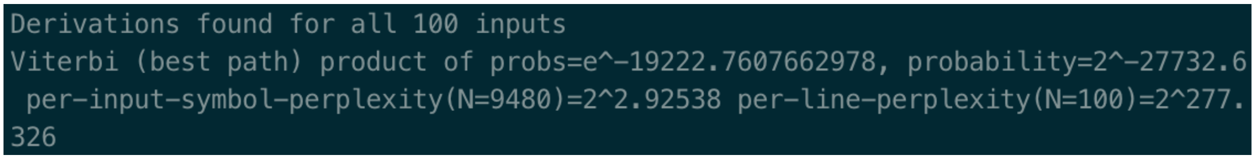
bigram:



```
Derivations found for all 100 inputs
Viterbi (best path) product of probs=e^-22539.9452505938, probability=2^-32518.3
 per-input-symbol-perplexity(N=9480)=2^3.4302 per-line-perplexity(N=100)=2^325.1
83
```

entropy: 3.4302

trigram:

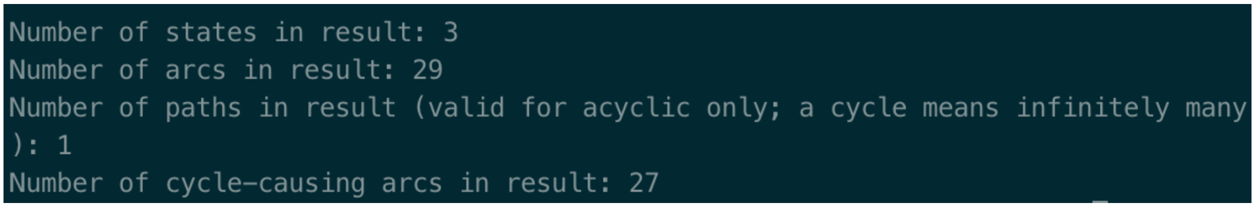


```
Derivations found for all 100 inputs
Viterbi (best path) product of probs=e^-19222.7607662978, probability=2^-27732.6
 per-input-symbol-perplexity(N=9480)=2^2.92538 per-line-perplexity(N=100)=2^277.
326
```

entropy: 2.92538

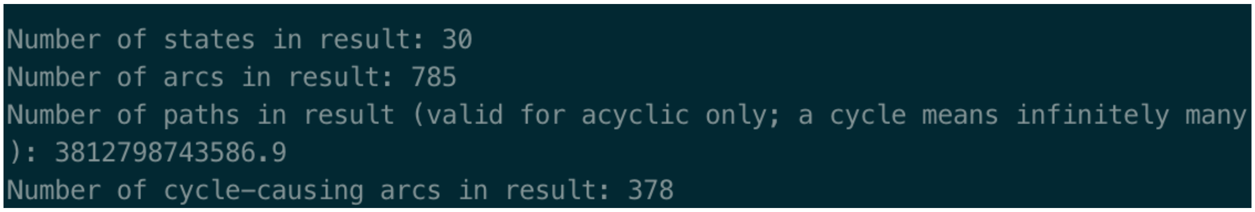
3.

unigram:



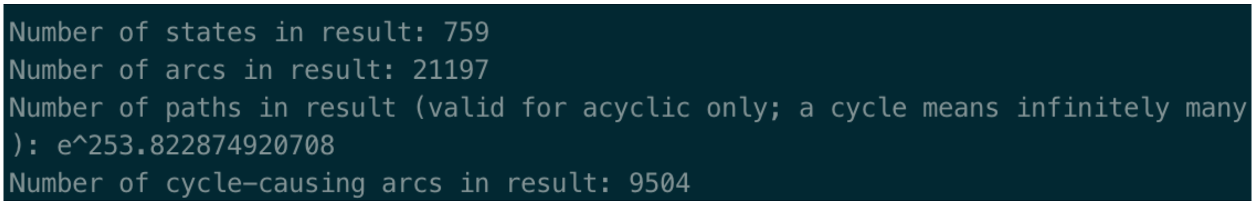
```
Number of states in result: 3
Number of arcs in result: 29
Number of paths in result (valid for acyclic only; a cycle means infinitely many
): 1
Number of cycle-causing arcs in result: 27
```

bigram:



```
Number of states in result: 30
Number of arcs in result: 785
Number of paths in result (valid for acyclic only; a cycle means infinitely many
): 3812798743586.9
Number of cycle-causing arcs in result: 378
```

trigram:



```
Number of states in result: 759
Number of arcs in result: 21197
Number of paths in result (valid for acyclic only; a cycle means infinitely many
): e^253.822874920708
Number of cycle-causing arcs in result: 9504
```

4.

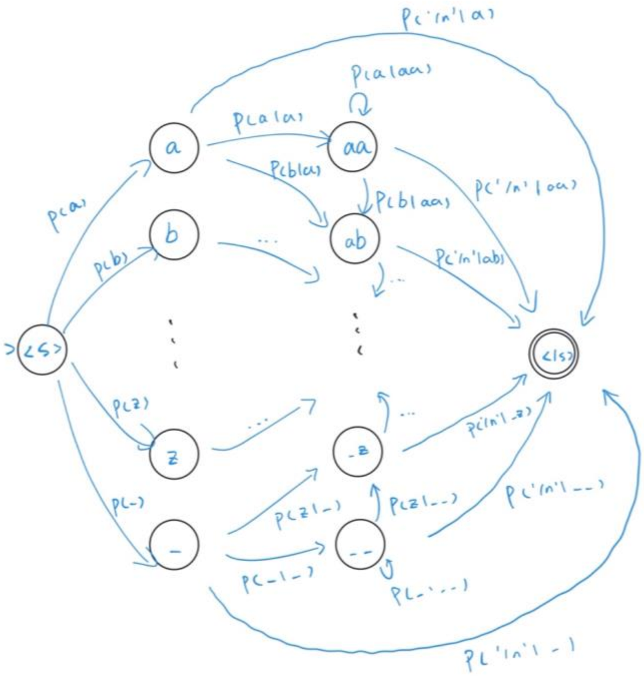
I used the Witten-Bell smoothing.

$$P(w_i|w_{i-1} = \lambda_{w_{i-1}} P_{ML}(w_i|w_{i-1} + \left(1 - \lambda_{w_{i-1}}\right) P(w_i)$$

$$\lambda_{w_{i-1}} = 1 - \frac{T(w_{i-1})}{T(w_{i-1}) + N(w_{i-1})}$$

I used ‘train.txt’ for training, ‘dev.txt’ for fiddling λ , and ‘test.txt’ for testing.

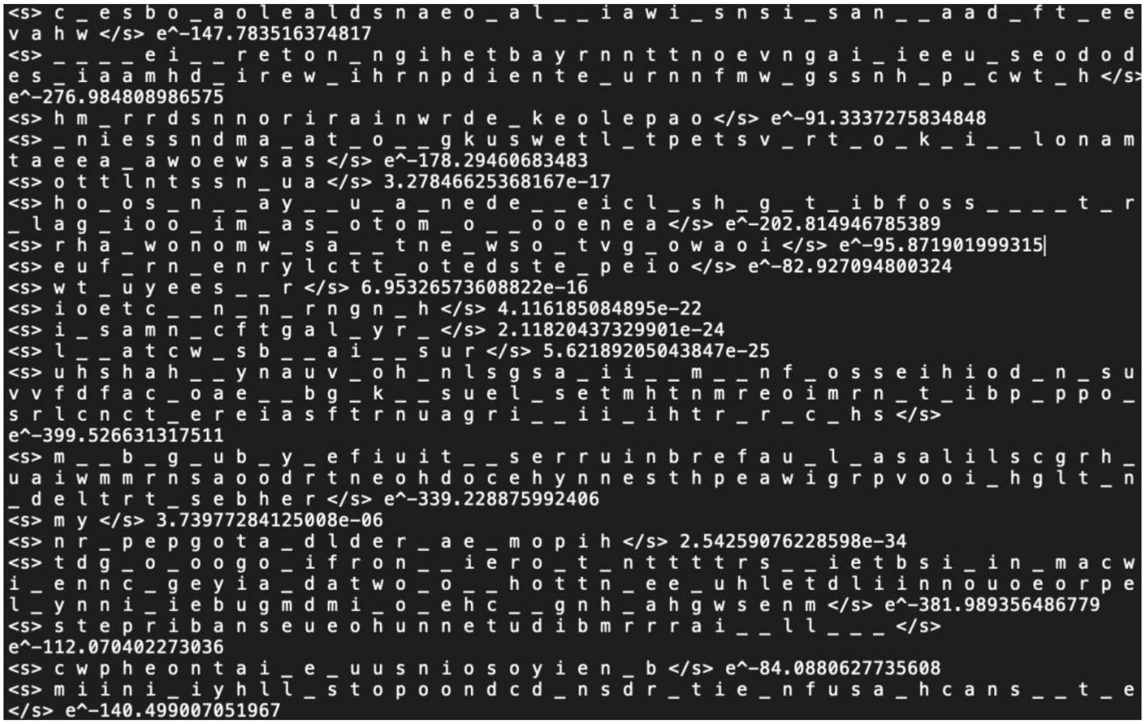
5.



2. Using Language Models

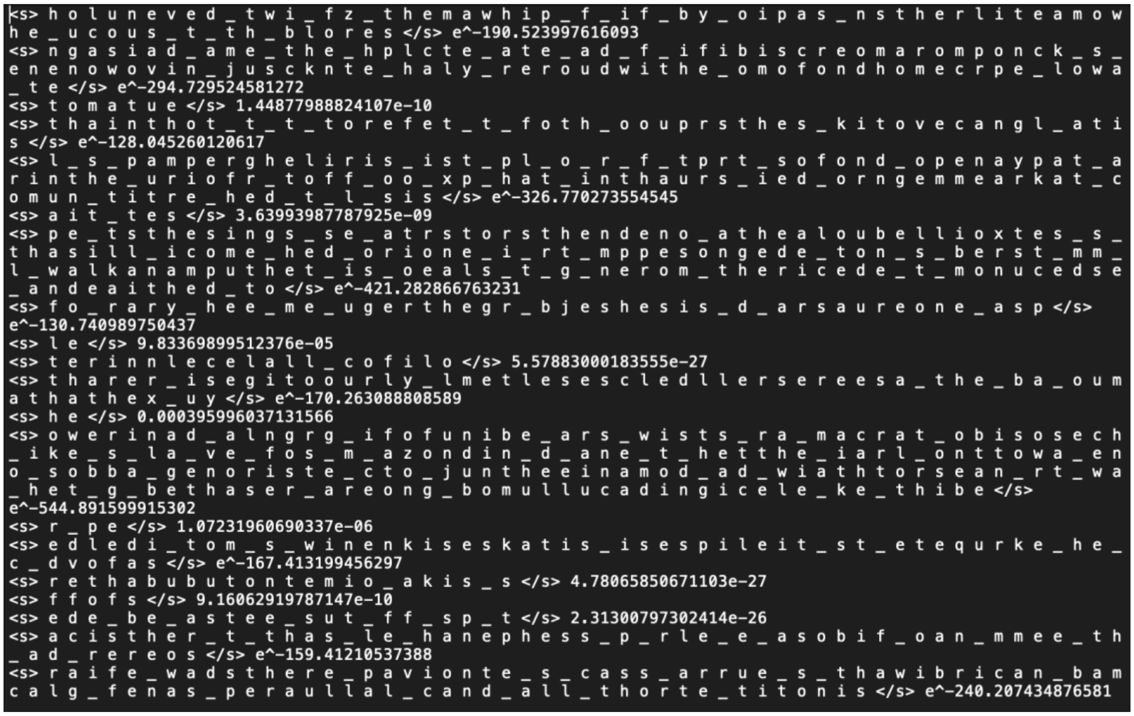
1.

unigram: doesn’t make sense.



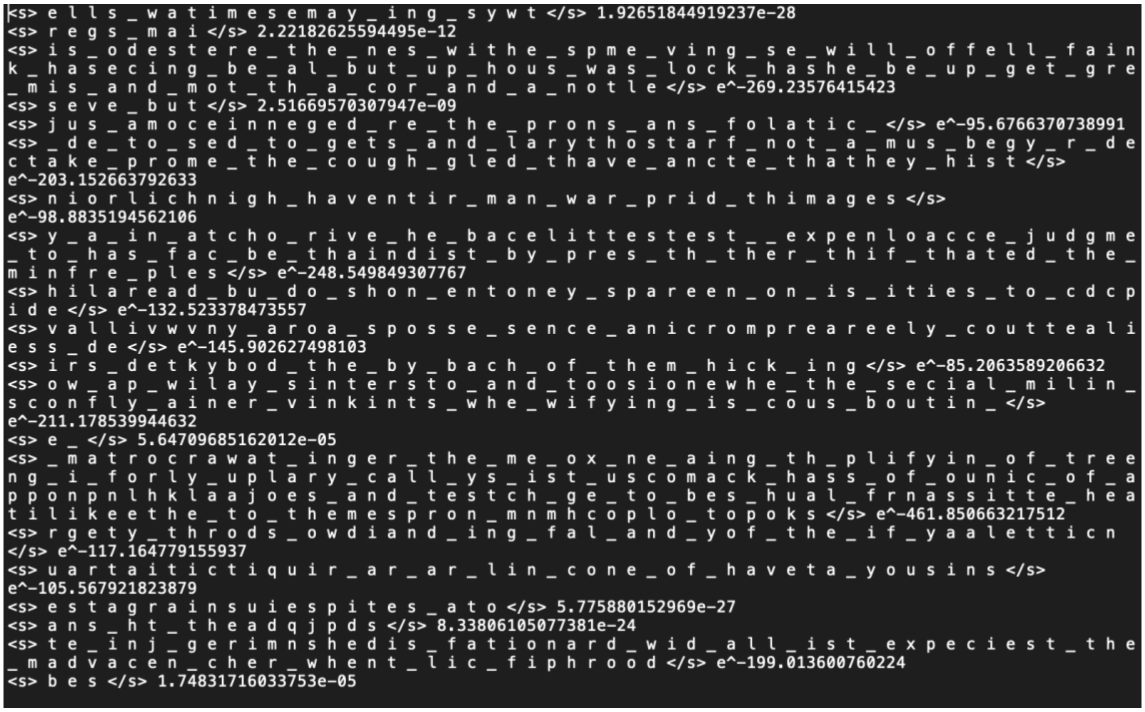
```
<s> c _ e s b o _ a o l e a l d s n a e o _ a l _ _ i a w i _ s n s i _ s a n _ _ a a d _ f t _ e e
v a h w </s> e^-147.783516374817
<s> _ _ _ _ e i _ _ r e t o n _ n g i h e t b a y r n n t t n o e v n g a i _ i e e u _ s e o d o d
e s _ i a a m h d _ i r e w _ i h r n p d i e n t e _ u r n n f m w _ g s s n h _ p _ c w t _ h </s>
e^-276.984808986575
<s> h m _ r r d s n n o r i r a i n w r d e _ k e o l e p a o </s> e^-91.3337275834848
<s> _ n i e s s n d m a _ a t _ o _ _ g k u s w e t l _ t p e t s v _ r t _ o _ k _ i _ _ l o n a m
t a e e a _ a w o e w s a s </s> e^-178.29460683483
<s> o t t l n t s s n _ u a </s> 3.27846625368167e-17
<s> h o _ o s _ n _ _ a y _ _ u _ a _ n e d e _ _ e i c l _ s h _ g _ t _ i b f o s s _ _ _ _ t _ r
_ l a g _ i o o _ i m _ a s _ o t o m _ o _ _ o o e n e a </s> e^-202.814946785389
<s> r h a _ w o n o m w _ s a _ _ t n e _ w s o _ t v g _ o w a o i </s> e^-95.871901999315
<s> e u f _ r n _ e n r y l c t t _ o t e d s t e _ p e i o </s> e^-82.927094800324
<s> w t _ u y e e s _ _ r </s> 6.95326573608822e-16
<s> i o e t c _ _ n _ n _ r n g n _ h </s> 4.116185084895e-22
<s> i _ s a m n _ c f t g a l _ y r _ </s> 2.11820437329901e-24
<s> l _ _ a t c w _ s b _ _ a i _ _ s u r </s> 5.62189205043847e-25
<s> u h s h a h _ _ y n a u v _ o h _ n l s g s a _ i i _ _ m _ _ n f _ o s s e i h i o d _ n _ s u
v v f d f a c _ o a e _ _ b g _ k _ _ s u e l _ s e t m h t n m r e o i m r n _ t _ i b p _ p p o _
s r l c n c t _ e r e i a s f t r n u a g r i _ _ i i _ i h t r _ r _ c _ h s </s>
e^-399.526631317511
<s> m _ _ b _ g _ u b _ y _ e f i u i t _ _ s e r r u i n b r e f a u _ l _ a s a l i l s c g r h _
u a i w m m r n s a o o d r t n e o h d o c e h y n n e s t h p e a w i g r p v o o i _ h g l t _ n
_ d e l t r t _ s e b h e r </s> e^-339.228875992406
<s> m y </s> 3.73977284125008e-06
<s> n r _ p e p g o t a _ d l d e r _ a e _ m o p i h </s> 2.54259076228598e-34
<s> t d g _ o _ o o g o _ i f r o n _ _ i e r o _ t _ n t t t t r s _ _ i e t b s i _ i n _ m a c w
i _ e n n c _ g e y i a _ d a t w o _ o _ _ h o t t n _ e e _ u h l e t d l i i n n o u o e o r p e
l _ y n n i _ i e b u g m d m i _ o _ e h c _ _ g n h _ a h g w s e n m </s> e^-381.989356486779
<s> s t e p r i b a n s e u e o h u n n e t u d i b m r r r a i _ _ l l _ _ _ </s>
e^-112.070402273036
<s> c w p h e o n t a i _ e _ u u s n i o s o y i e n _ b </s> e^-84.0880627735608
<s> m i i n i _ i y h l l _ s t o p o o n d c d _ n s d r _ t i e _ n f u s a _ h c a n s _ _ t _ e
</s> e^-140.499007051967
```

bigram: has correctly spelled a few words (me, be, the)



```
<s> h o l u n e v e d _ t w i _ f z _ t h e m a w h i p _ f _ i f _ b y _ o i p a s _ n s t h e r l i t e a m o w
h e _ u c o u s _ t _ t h _ b l o r e s </s> e^-190.523997616093
<s> n g a s i a d _ a m e _ t h e _ h p l c t e _ a t e _ a d _ f _ i f i b i s c r e o m a r o m p o n c k _ s _
e n e n o w o v i n _ j u s c k n t e _ h a l y _ r e r o u d w i t h e _ o m o f o n d h o m e c r p e _ l o w a
_ t e </s> e^-294.729524581272
<s> t o m a t u e </s> 1.44877988824107e-10
<s> t h a i n t h o t _ t _ t _ t o r e f e t _ t _ f o t h _ o o u p r s t h e s _ k i t o v e c a n g l _ a t i
s </s> e^-128.045260120617
<s> l _ s _ p a m p e r g h e l i r i s _ i s t _ p l _ o _ r _ f _ t p r t _ s o f o n d _ o p e n a y p a t _ a
r i n t h e _ u r i o f r _ t o f f _ o o _ x p _ h a t _ i n t h a u r s _ i e d _ o r n g e m m e a r k a t _ c
o m u n _ t i t r e _ h e d _ t _ l _ s i s </s> e^-326.770273554545
<s> a i t _ t e s </s> 3.63993987787925e-09
<s> p e _ t s t h e s i n g s _ s e _ a t r s t o r s t h e n d e n o _ a t h e a l o u b e l l i o x t e s _ s _
t h a s i l l _ i c o m e _ h e d _ o r i o n e _ i _ r t _ m p p e s o n g e d e _ t o n _ s _ b e r s t _ m m _
l _ w a l k a n a m p u t h e t _ i s _ o e a l s _ t _ g _ n e r o m _ t h e r i c e d e _ t _ m o n u c e d s e
_ a n d e a i t h e d _ t o </s> e^-421.282866763231
<s> f o _ r a r y _ h e e _ m e _ u g e r t h e g r _ b j e s h e s i s _ d _ a r s a u r e o n e _ a s p </s>
e^-130.740989750437
<s> l e </s> 9.83369899512376e-05
<s> t e r i n n l e c e l a l l _ c o f i l o </s> 5.57883000183555e-27
<s> t h a r e r _ i s e g i t o o u r l y _ l m e t l e s e s c l e d l l e r s e r e e s a _ t h e _ b a _ o u m
a t h a t h e x _ u y </s> e^-170.263088808589
<s> h e </s> 0.000395996037131566
<s> o w e r i n a d _ a l n g r g _ i f o f u n i b e _ a r s _ w i s t s _ r a _ m a c r a t _ o b i s o s e c h
_ i k e _ s _ l a _ v e _ f o s _ m _ a z o n d i n _ d _ a n e _ t _ h e t t h e _ i a r l _ o n t t o w a _ e n
o _ s o b b a _ g e n o r i s t e _ c t o _ j u n t h e e i n a m o d _ a d _ w i a t h t o r s e a n _ r t _ w a
_ h e t _ g _ b e t h a s e r _ a r e o n g _ b o m u l l u c a d i n g i c e l e _ k e _ t h i b e </s>
e^-544.891599915302
<s> r _ p e </s> 1.07231960690337e-06
<s> e d l e d i _ t o m _ s _ w i n e n k i s e s k a t i s _ i s e s p i l e i t _ s t _ e t e q u r k e _ h e _
c _ d v o f a s </s> e^-167.413199456297
<s> r e t h a b u b u t o n t e m i o _ a k i s _ s </s> 4.78065850671103e-27
<s> f f o f s </s> 9.16062919787147e-10
<s> e d e _ b e _ a s t e e _ s u t _ f f _ s p _ t </s> 2.31300797302414e-26
<s> a c i s t h e r _ t _ t h a s _ l e _ h a n e p h e s s _ p _ r l e _ e _ a s o b i f _ o a n _ m m e e _ t h
_ a d _ r e r e o s </s> e^-159.41210537388
<s> r a i f e _ w a d s t h e r e _ p a v i o n t e _ s _ c a s s _ a r r u e _ s _ t h a w i b r i c a n _ b a m
c a l g _ f e n a s _ p e r a u l l a l _ c a n d _ a l l _ t h o r t e _ t i t o n i s </s> e^-240.207434876581
```

trigram: makes more sense, appears more correct words (them, hick, man)



```
<s> e l l s _ w a t i m e s e m a y _ i n g _ s y w t </s> 1.92651844919237e-28
<s> r e g s _ m a i </s> 2.22182625594495e-12
<s> i s _ o d e s t e r e _ t h e _ n e s _ w i t h e _ s p m e _ v i n g _ s e _ w i l l _ o f f e l l _ f a i n
k _ h a s e c i n g _ b e _ a l _ b u t _ u p _ h o u s _ w a s _ l o c k _ h a s h e _ b e _ u p _ g e t _ g r e
_ m i s _ a n d _ m o t _ t h _ a _ c o r _ a n d _ a _ n o t l e </s> e^-269.23576415423
<s> s e v e _ b u t </s> 2.51669570307947e-09
<s> j u s _ a m o c e i n n e g e d _ r e _ t h e _ p r o n s _ a n s _ f o l a t i c _ </s> e^-95.6766370738991
<s> _ d e _ t o _ s e d _ t o _ g e t s _ a n d _ l a r y t h o s t a r f _ n o t _ a _ m u s _ b e g y _ r _ d e
c t a k e _ p r o m e _ t h e _ c o u g h _ g l e d _ t h a v e _ a n c t e _ t h a t h e y _ h i s t </s>
e^-203.152663792633
<s> n i o r l i c h n i g h _ h a v e n t i r _ m a n _ w a r _ p r i d _ t h i m a g e s </s>
e^-98.8835194562106
<s> y _ a _ i n _ a t c h o _ r i v e _ h e _ b a c e l i t t e s t e s t _ _ e x p e n l o a c c e _ j u d g m e
_ t o _ h a s _ f a c _ b e _ t h a i n d i s t _ b y _ p r e s _ t h _ t h e r _ t h i f _ t h a t e d _ t h e _
m i n f r e _ p l e s </s> e^-248.549849307767
<s> h i l a r e a d _ b u _ d o _ s h o n _ e n t o n e y _ s p a r e e n _ o n _ i s _ i t i e s _ t o _ c d c p
i d e </s> e^-132.523378473557
<s> v a l l i v w v n y _ a r o a _ s p o s s e _ s e n c e _ a n i c r o m p r e a r e e l y _ c o u t t e a l i
e s s _ d e </s> e^-145.902627498103
<s> i r s _ d e t k y b o d _ t h e _ b y _ b a c h _ o f _ t h e m _ h i c k _ i n g </s> e^-85.2063589206632
<s> o w _ a p _ w i l a y _ s i n t e r s t o _ a n d _ t o o s i o n e w h e _ t h e _ s e c i a l _ m i l i n _
s c o n f l y _ a i n e r _ v i n k i n t s _ w h e _ w i f y i n g _ i s _ c o u s _ b o u t i n _ </s>
e^-211.178539944632
<s> e _ </s> 5.64709685162012e-05
<s> _ m a t r o c r a w a t _ i n g e r _ t h e _ m e _ o x _ n e _ a i n g _ t h _ p l i f y i n _ o f _ t r e e
n g _ i _ f o r l y _ u p l a r y _ c a l l _ y s _ i s t _ u s c o m a c k _ h a s s _ o f _ o u n i c _ o f _ a
p p o n p n l h k l a a j o e s _ a n d _ t e s t c h _ g e _ t o _ b e s _ h u a l _ f r n a s s i t t e _ h e a
t i l i k e e t h e _ t o _ t h e m e s p r o n _ m n m h c o p l o _ t o p o k s </s> e^-461.850663217512
<s> r g e t y _ t h r o d s _ o w d i a n d _ i n g _ f a l _ a n d _ y o f _ t h e _ i f _ y a a l e t t i c n
</s> e^-117.164779155937
<s> u a r t a i t i c t i q u i r _ a r _ a r _ l i n _ c o n e _ o f _ h a v e t a _ y o u s i n s </s>
e^-105.567921823879
<s> e s t a g r a i n s u i e s p i t e s _ a t o </s> 5.775880152969e-27
<s> a n s _ h t _ t h e a d q j p d s </s> 8.33806105077381e-24
<s> t e _ i n j _ g e r i m n s h e d i s _ f a t i o n a r d _ w i d _ a l l _ i s t _ e x p e c i e s t _ t h e
_ m a d v a c e n _ c h e r _ w h e n t _ l i c _ f i p h r o o d </s> e^-199.013600760224
<s> b e s </s> 1.74831716033753e-05
```

2.

1. Use ‘sed -e 's/[aeiou]//g' test.txt > test.txt.novowels’ to remove vowels.

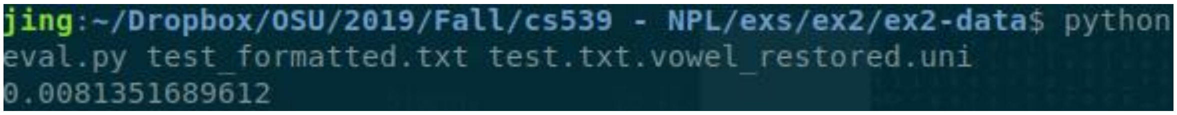
2. Use ‘carmel wfsa remove-vowels.fst > wfsa.fst’ to combine fsa and fst

3. Use ‘cat test.txt.novowels | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' | carmel -sribIEWk 1 wfsa.fst > test.txt.vowel_restored.{uni,bi,tri}’ to restore.

4. Use ‘cat test.txt | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' > test_formatted.txt’ to format the test.txt for evaluation.

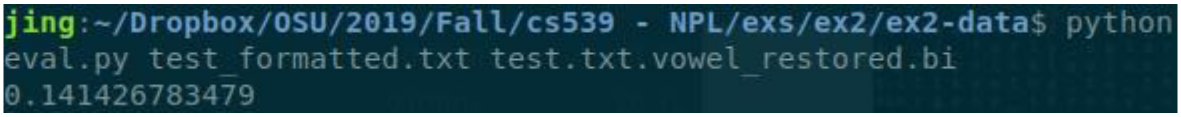
5. Use ‘python eval.py test_formatted.txt test.txt.vowel_restored.uni’ to get the accuracy.

Unigram:



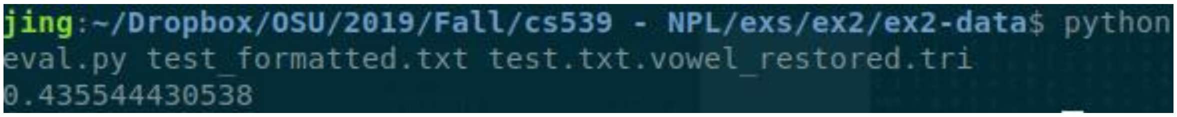
```
jing:~/Dropbox/OSU/2019/Fall/cs539 - NPL/exs/ex2/ex2-data$ python
eval.py test_formatted.txt test.txt.vowel_restored.uni
0.0081351689612
```

Bigram:



```
jing:~/Dropbox/OSU/2019/Fall/cs539 - NPL/exs/ex2/ex2-data$ python
eval.py test_formatted.txt test.txt.vowel_restored.bi
0.141426783479
```

Trigram:



```
jing:~/Dropbox/OSU/2019/Fall/cs539 - NPL/exs/ex2/ex2-data$ python
eval.py test_formatted.txt test.txt.vowel_restored.tri
0.435544430538
```

3.

1. Use ‘sed -e 's/[ ]//g' test.txt > test.txt.nospaces’ to remove spaces.

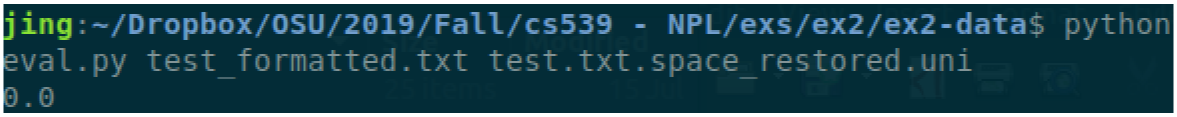
2. Use ‘carmel wfsa remove-spaces.fst > wfsa.fst’ to combine fsa and fst

3. Use ‘cat test.txt.nospaces | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' | carmel -sribIEWk 1 wfsa.fst > test.txt.space_restored.{uni,bi,tri}’ to restore.

4. Use ‘cat test.txt | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' > test_formatted.txt’ to format the test.txt for evaluation.

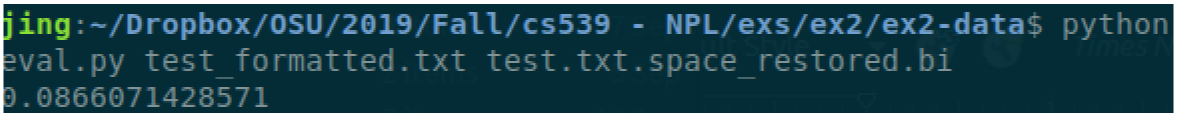
5. Use ‘python eval.py test_formatted.txt test.txt.space_restored.uni’ to get the accuracy.

Unigram:



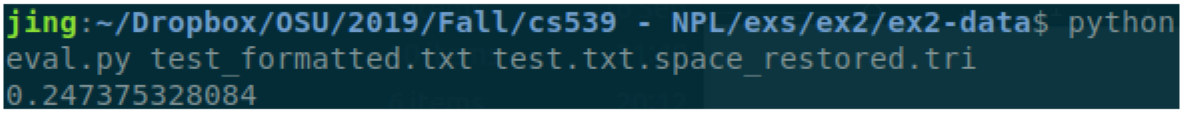
```
jing:~/Dropbox/OSU/2019/Fall/cs539 - NPL/exs/ex2/ex2-data$ python
eval.py test_formatted.txt test.txt.space_restored.uni
0.0
```

Bigram:



```
jing:~/Dropbox/OSU/2019/Fall/cs539 - NPL/exs/ex2/ex2-data$ python
eval.py test_formatted.txt test.txt.space_restored.bi
0.0866071428571
```

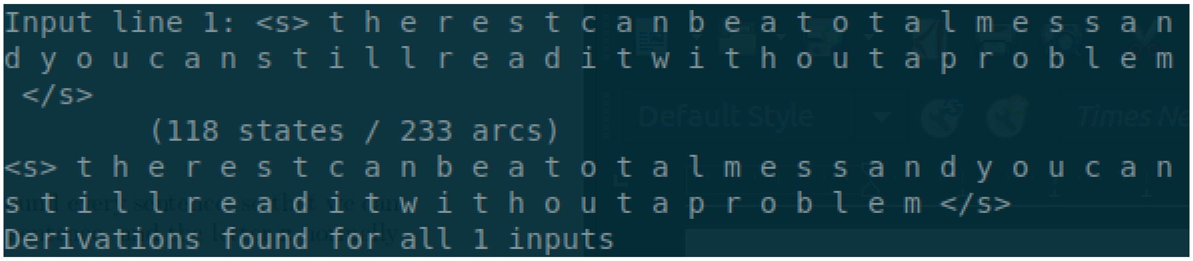
Trigram:



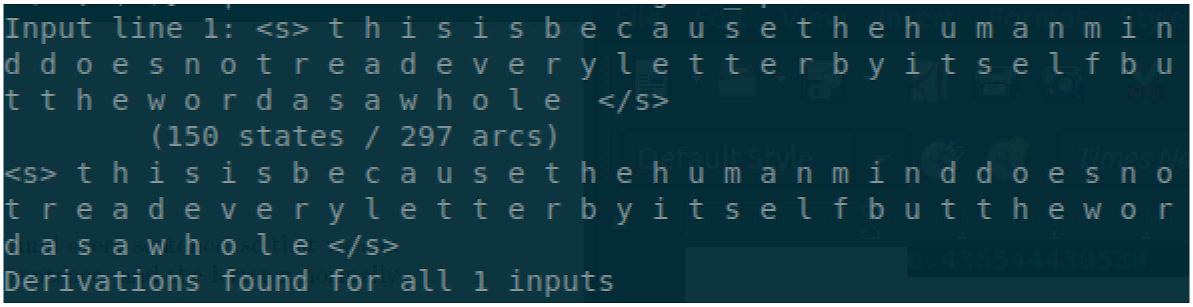
```
jing:~/Dropbox/OSU/2019/Fall/cs539 - NPL/exs/ex2/ex2-data$ python
eval.py test_formatted.txt test.txt.space_restored.tri
0.247375328084
```

Sentences:

Unigram:

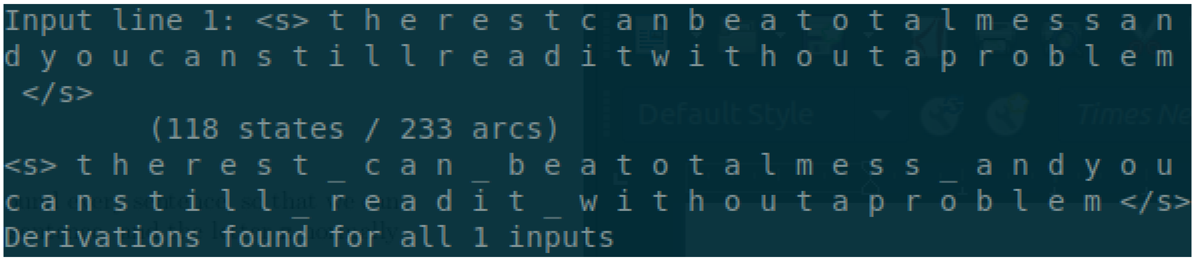


```
Input line 1: <s> t h e r e s t c a n b e a t o t a l m e s s a n
d y o u c a n s t i l l r e a d i t w i t h o u t a p r o b l e m
 </s>
        (118 states / 233 arcs)
<s> t h e r e s t c a n b e a t o t a l m e s s a n d y o u c a n
s t i l l r e a d i t w i t h o u t a p r o b l e m </s>
Derivations found for all 1 inputs
```

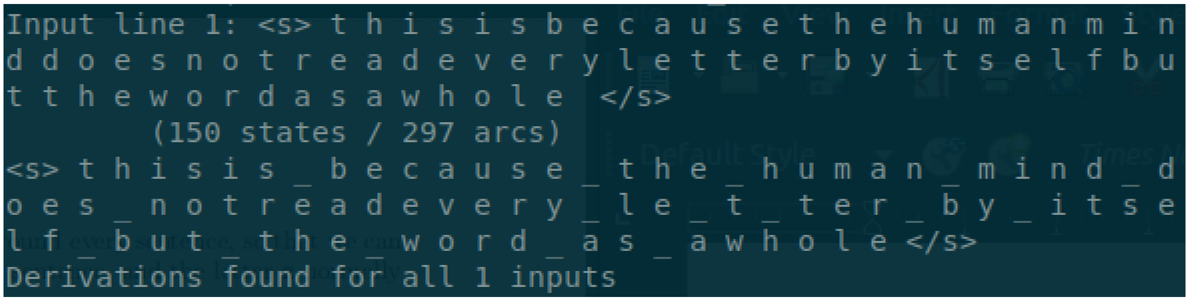


```
Input line 1: <s> t h i s i s b e c a u s e t h e h u m a n m i n
d d o e s n o t r e a d e v e r y l e t t e r b y i t s e l f b u
t t h e w o r d a s a w h o l e  </s>
        (150 states / 297 arcs)
<s> t h i s i s b e c a u s e t h e h u m a n m i n d d o e s n o
t r e a d e v e r y l e t t e r b y i t s e l f b u t t h e w o r
d a s a w h o l e </s>
Derivations found for all 1 inputs
```

Bigram:

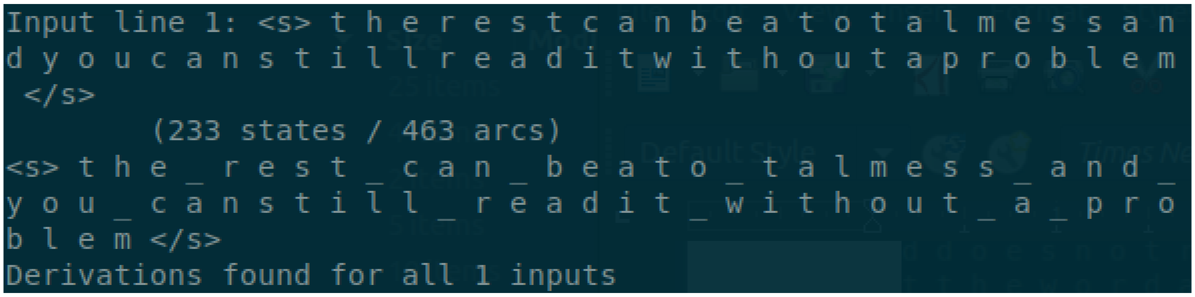


```
Input line 1: <s> t h e r e s t c a n b e a t o t a l m e s s a n
d y o u c a n s t i l l r e a d i t w i t h o u t a p r o b l e m
 </s>
        (118 states / 233 arcs)
<s> t h e r e s t _ c a n _ b e a t o t a l m e s s _ a n d y o u
c a n s t i l l _ r e a d i t _ w i t h o u t a p r o b l e m </s>
Derivations found for all 1 inputs
```

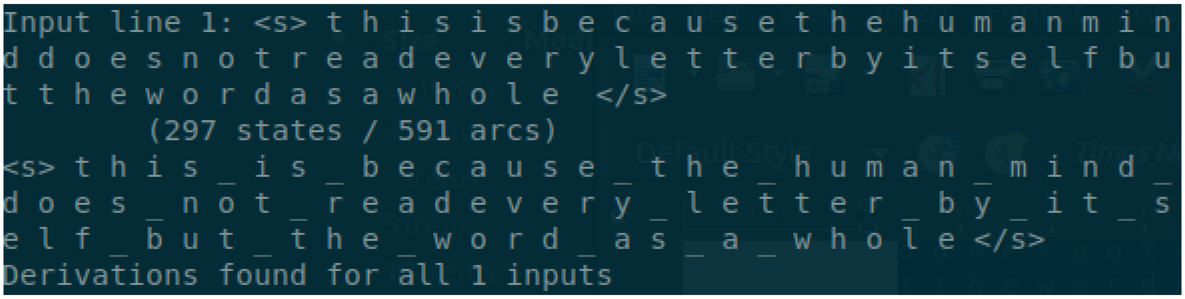


```
Input line 1: <s> t h i s i s b e c a u s e t h e h u m a n m i n
d d o e s n o t r e a d e v e r y l e t t e r b y i t s e l f b u
t t h e w o r d a s a w h o l e  </s>
        (150 states / 297 arcs)
<s> t h i s i s _ b e c a u s e _ t h e _ h u m a n _ m i n d _ d
o e s _ n o t r e a d e v e r y _ l e _ t _ t e r _ b y _ i t s e
l f _ b u t _ t h e _ w o r d _ a s _ a w h o l e </s>
Derivations found for all 1 inputs
```

Trigram:



```
Input line 1: <s> t h e r e s t c a n b e a t o t a l m e s s a n
d y o u c a n s t i l l r e a d i t w i t h o u t a p r o b l e m
 </s>
        (233 states / 463 arcs)
<s> t h e _ r e s t _ c a n _ b e a t o _ t a l m e s s _ a n d _
y o u _ c a n s t i l l _ r e a d i t _ w i t h o u t _ a _ p r o
b l e m </s>
Derivations found for all 1 inputs
```



```
Input line 1: <s> t h i s i s b e c a u s e t h e h u m a n m i n
d d o e s n o t r e a d e v e r y l e t t e r b y i t s e l f b u
t t h e w o r d a s a w h o l e  </s>
        (297 states / 591 arcs)
<s> t h i s _ i s _ b e c a u s e _ t h e _ h u m a n _ m i n d _
d o e s _ n o t _ r e a d e v e r y _ l e t t e r _ b y _ i t _ s
e l f _ b u t _ t h e _ w o r d _ a s _ a _ w h o l e </s>
Derivations found for all 1 inputs
```

4. From the observations in these experiments, the accuracy results of restoring vowels are better than the results of restoring space. In addition, the number of states and arcs of generated vowels.fst are greater than the generated space.fst's. In other words, the restoring of vowels has more conditions to judge than the restoring of space. As a result, restoring vowels is easier.