# EX 2: Language Models and Entropy

Jing Huang (933039277)

## 1. Training Language Models

2.

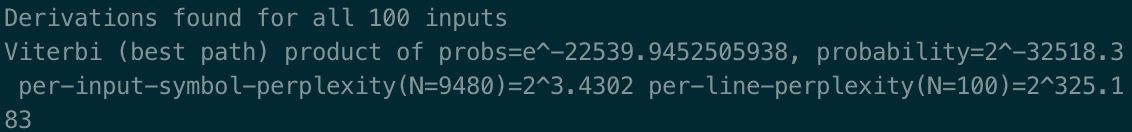
unigram:

A close up of a sign

Description automatically generated

entropy: 4.11494

bigram:



entropy: 3.4302

trigram:

A close up of a sign

Description automatically generated

entropy: 2.92538

3.

unigram:

A close up of a sign

Description automatically generated

bigram:

A close up of a sign

Description automatically generated

trigram:

A close up of a sign

Description automatically generated

4.

I used the Witten-Bell smoothing.

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

5.

A close up of a logo

Description automatically generated

# 2. Using Language Models

1.

unigram: doesn’t make sense.

A screenshot of a cell phone

Description automatically generated

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

A close up of text on a black background

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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:

 Bigram:



Trigram:



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:

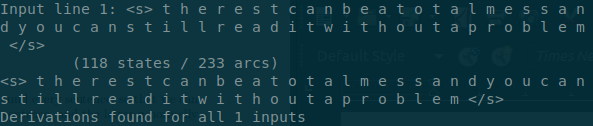
 Bigram:

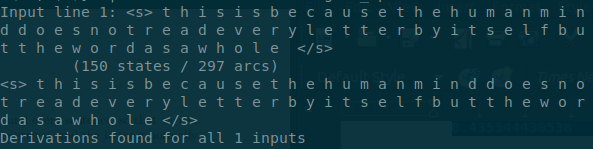
 Trigram:



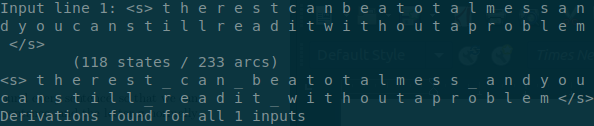
Sentences:

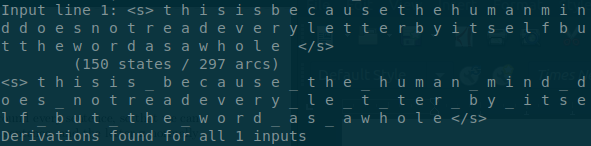
Unigram:



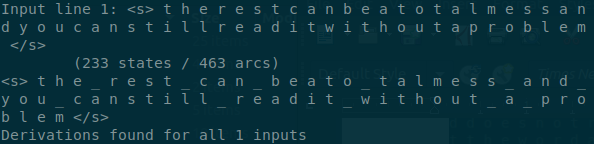


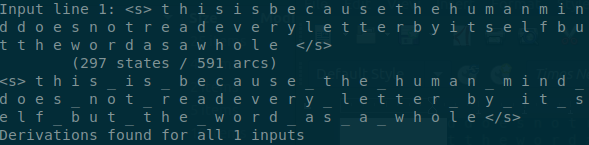
Bigram:





Trigram:





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.