Language modelling using SRILM

Aman kumar 17025

toy dataset

#Using a bigram language model based on the training data, estimate probability of below sentences

(a) S: **Sam I do I like** (b) S: **Sam I am**

In [ ]:

```
#printing training file
!cat sample-data/sample-train.txt
```

```
I am Sam
Sam I am
Sam I like
Sam I do like
do I like Sam
```

In [ ]:

```
#printing test file 1
!cat sample-data/test-q1.txt
```

```
Sam I do I like
Sam I am
```

In [ ]:

```
#printing test file 2
!cat sample-data/test-q5.txt
```

```
Sam I do like linguistics
```

In [ ]:

```
# for permission for binary files
!chmod +x ngram
```

In [ ]:

```
!chmod +x ngram-count
```

In [ ]:

```
#Training a bigram language model using training data

!./ngram-count -text sample-data/sample-train.txt -order 2 -write sample-data/files/sample-
-lm sample-data/files/sample-train-bigram.lm -addsmooth 0
```

In [ ]:

```
!cat sample-data/files/sample-train-bigram.count
```

```
<s>     5
<s> I   1
<s> Sam 3
<s> do  1
I       5
I am    2
I like  2
I do    1
am      2
am Sam  1
am </s> 1
Sam     5
Sam </s>        2
Sam I   3
</s>    5
like    3
like </s>       2
like Sam        1
do      2
do like 1
do I    1
```

In [ ]:

```
!cat sample-data/files/sample-train-bigram.lm
```

```
\data\
ngram 1=7
ngram 2=14

\1-grams:
-0.6434527      </s>
-99     <s>     -7.292433
-0.6434527      I       -7.866736
-0.6434527      Sam     -7.737127
-1.041393       am      -7.21857
-1.041393       do      -7.285517
-0.8653014      like    -99

\2-grams:
-0.69897        <s> I
-0.2218488      <s> Sam
-0.69897        <s> do
-0.39794        I am
-0.69897        I do
-0.39794        I like
-0.39794        Sam </s>
-0.2218488      Sam I
-0.30103        am </s>
-0.30103        am Sam
-0.30103        do I
-0.30103        do like
-0.1760913      like </s>
-0.4771213      like Sam

\end\
```

In [ ]:

```
#Testing the above bigram language model

!./ngram -lm sample-data/files/sample-train-bigram.lm -ppl sample-data/test-q1.txt -debug 2
```

```
reading 7 1-grams
reading 14 2-grams
Sam I do I like
        p( Sam | <s> )  = [2gram] 0.6 [ -0.221849 ]
        p( I | Sam ...)         = [2gram] 0.6 [ -0.221849 ]
        p( do | I ...)  = [2gram] 0.2 [ -0.69897 ]
        p( I | do ...)  = [2gram] 0.5 [ -0.30103 ]
        p( like | I ...)        = [2gram] 0.4 [ -0.39794 ]
        p( </s> | like ...)     = [2gram] 0.666667 [ -0.176091 ]
1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -2.01773 ppl= 2.16914 ppl1= 2.53248

Sam I am
        p( Sam | <s> )  = [2gram] 0.6 [ -0.221849 ]
        p( I | Sam ...)         = [2gram] 0.6 [ -0.221849 ]
        p( am | I ...)  = [2gram] 0.4 [ -0.39794 ]
        p( </s> | am ...)       = [2gram] 0.5 [ -0.30103 ]
1 sentences, 3 words, 0 OOVs
0 zeroprobs, logprob= -1.14267 ppl= 1.93049 ppl1= 2.40375

file sample-data/test-q1.txt: 2 sentences, 8 words, 0 OOVs
0 zeroprobs, logprob= -3.1604 ppl= 2.07033 ppl1= 2.48342
```

Using a unigram language model based on the training data, estimate probability of below sentences

(a) S: **Sam I do I like** (b) S: **Sam I am**

In [ ]:

```
!./ngram-count -text sample-data/sample-train.txt -order 1 -write sample-data/files/sample-
-lm sample-data/files/sample-train-unigram.lm -addsmooth 0
```

In [ ]:

```
!cat sample-data/files/sample-train-unigram.count
```

```
<s>     5
I       5
am      2
Sam     5
</s>    5
like    3
do      2
```

In [ ]:

```
!cat sample-data/files/sample-train-unigram.lm
```

```
\data\
ngram 1=7

\1-grams:
-0.6434527      </s>
-99     <s>
-0.6434527      I
-0.6434527      Sam
-1.041393       am
-1.041393       do
-0.8653014      like

\end\
```

In [ ]:

```
#Testing the unigram language model
!./ngram -lm sample-data/files/sample-train-unigram.lm -ppl sample-data/test-q1.txt -debug
```

```
reading 7 1-grams
Sam I do I like
        p( Sam | <s> )  = [1gram] 0.227273 [ -0.643453 ]
        p( I | Sam ...)         = [1gram] 0.227273 [ -0.643453 ]
        p( do | I ...)  = [1gram] 0.090909 [ -1.04139 ]
        p( I | do ...)  = [1gram] 0.227273 [ -0.643453 ]
        p( like | I ...)        = [1gram] 0.136364 [ -0.865301 ]
        p( </s> | like ...)     = [1gram] 0.227273 [ -0.643453 ]
1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -4.48051 ppl= 5.5815 ppl1= 7.87229

Sam I am
        p( Sam | <s> )  = [1gram] 0.227273 [ -0.643453 ]
        p( I | Sam ...)         = [1gram] 0.227273 [ -0.643453 ]
        p( am | I ...)  = [1gram] 0.090909 [ -1.04139 ]
        p( </s> | am ...)       = [1gram] 0.227273 [ -0.643453 ]
1 sentences, 3 words, 0 OOVs
0 zeroprobs, logprob= -2.97175 ppl= 5.53271 ppl1= 9.78552

file sample-data/test-q1.txt: 2 sentences, 8 words, 0 OOVs
0 zeroprobs, logprob= -7.45226 ppl= 5.56193 ppl1= 8.54146
```

Now use a unigram/bigram LM with above training data and estimate per word probability of the sentence below:

S: **Sam I do like linguistics**

In [ ]:

```
!cat sample-data/sample-train.txt
```

```
I am Sam
Sam I am
Sam I like
Sam I do like
do I like Sam
```

In [ ]:

```
!cat sample-data/test-q5.txt
```

```
Sam I do like linguistics
```

In [ ]:

```
#Testing the above bigram language model

!./ngram -lm sample-data/files/sample-train-bigram.lm -ppl sample-data/test-q5.txt -debug 2
```

```
reading 7 1-grams
reading 14 2-grams
Sam I do like linguistics
        p( Sam | <s> )  = [2gram] 0.6 [ -0.221849 ]
        p( I | Sam ...)         = [2gram] 0.6 [ -0.221849 ]
        p( do | I ...)  = [2gram] 0.2 [ -0.69897 ]
        p( like | do ...)       = [2gram] 0.5 [ -0.30103 ]
        p( <unk> | like ...)    = [OOV] 0 [ -inf ]
        p( </s> | <unk> ...)    = [1gram] 0.227273 [ -0.643453 ]
1 sentences, 5 words, 1 OOVs
0 zeroprobs, logprob= -2.08715 ppl= 2.61475 ppl1= 3.32497

file sample-data/test-q5.txt: 1 sentences, 5 words, 1 OOVs
0 zeroprobs, logprob= -2.08715 ppl= 2.61475 ppl1= 3.32497
```

In [ ]:

```
#Testing the above unigram language model

!./ngram -lm sample-data/files/sample-train-unigram.lm -ppl sample-data/test-q5.txt -debug
```

```
reading 7 1-grams
Sam I do like linguistics
        p( Sam | <s> )  = [1gram] 0.227273 [ -0.643453 ]
        p( I | Sam ...)         = [1gram] 0.227273 [ -0.643453 ]
        p( do | I ...)  = [1gram] 0.090909 [ -1.04139 ]
        p( like | do ...)       = [1gram] 0.136364 [ -0.865301 ]
        p( <unk> | like ...)    = [OOV] 0 [ -inf ]
        p( </s> | <unk> ...)    = [1gram] 0.227273 [ -0.643453 ]
1 sentences, 5 words, 1 OOVs
0 zeroprobs, logprob= -3.83705 ppl= 5.85343 ppl1= 9.10465

file sample-data/test-q5.txt: 1 sentences, 5 words, 1 OOVs
0 zeroprobs, logprob= -3.83705 ppl= 5.85343 ppl1= 9.10465
```

Applying Laplace Smoothing to rescue

#

- unigram

In [ ]:

```
#Training a laplace smoothed unigram language model using training data

!./ngram-count -text sample-data/sample-train.txt -order 1 -write sample-data/files/sample-
-lm sample-data/files/sample-train-unigram-smoothed.lm -addsmooth 1 -unk

#Testing the above unigram language model

!./ngram -lm sample-data/files/sample-train-unigram-smoothed.lm -ppl sample-data/test-q5.tx
```

```
reading 8 1-grams
sample-data/files/sample-train-unigram-smoothed.lm: line 8: warning: non-zer
o probability for <unk> in closed-vocabulary LM
Sam I do like linguistics
        p( Sam | <s> )  = [1gram] 0.206897 [ -0.684247 ]
        p( I | Sam ...)         = [1gram] 0.206897 [ -0.684247 ]
        p( do | I ...)  = [1gram] 0.103448 [ -0.985277 ]
        p( like | do ...)       = [1gram] 0.137931 [ -0.860338 ]
        p( <unk> | like ...)    = [1gram] 0.0344828 [ -1.4624 ]
        p( </s> | <unk> ...)    = [1gram] 0.206897 [ -0.684247 ]
1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -5.36075 ppl= 7.82454 ppl1= 11.8073

file sample-data/test-q5.txt: 1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -5.36075 ppl= 7.82454 ppl1= 11.8073
```

- bigram

In [ ]:

```
#Training a laplace smoothed bigram language model using training data

!./ngram-count -text sample-data/sample-train.txt -order 2 -write sample-data/files/sample-
-lm sample-data/files/sample-train-bigram-smoothed.lm -addsmooth 1 -unk

#Testing the above bigram language model

!./ngram -lm sample-data/files/sample-train-bigram-smoothed.lm -ppl sample-data/test-q5.txt
```

```
reading 8 1-grams
sample-data/files/sample-train-bigram-smoothed.lm: line 9: warning: non-zero
probability for <unk> in closed-vocabulary LM
reading 14 2-grams
Sam I do like linguistics
        p( Sam | <s> )  = [2gram] 0.333333 [ -0.477121 ]
        p( I | Sam ...)         = [2gram] 0.333333 [ -0.477121 ]
        p( do | I ...)  = [2gram] 0.166667 [ -0.778151 ]
        p( like | do ...)       = [2gram] 0.222222 [ -0.653212 ]
        p( <unk> | like ...)    = [1gram] 0.0294118 [ -1.53148 ]
        p( </s> | <unk> ...)    = [1gram] 0.206897 [ -0.684247 ]
1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -4.60133 ppl= 5.8464 ppl1= 8.32274

file sample-data/test-q5.txt: 1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -4.60133 ppl= 5.8464 ppl1= 8.32274
```

#Linear Interpolation of Unigram and Bigram LM

In [ ]:

```
!cat sample-data/test-q5.txt
```

```
Sam I do like linguistics
```

In [ ]:

```
!./ngram -lm sample-data/files/sample-train-bigram-smoothed.lm -mix-lm sample-data/files/sa
-lambda 0.5 -ppl sample-data/test-q5.txt -debug 2
```

```
reading 8 1-grams
sample-data/files/sample-train-bigram-smoothed.lm: line 9: warning: non-zero
probability for <unk> in closed-vocabulary LM
reading 14 2-grams
reading 8 1-grams
sample-data/files/sample-train-unigram-smoothed.lm: line 8: warning: non-zer
o probability for <unk> in closed-vocabulary LM
Sam I do like linguistics
        p( Sam | <s> )  = [2gram] 0.270115 [ -0.568451 ]
        p( I | Sam ...)         = [2gram] 0.270115 [ -0.568451 ]
        p( do | I ...)  = [2gram] 0.135057 [ -0.869481 ]
        p( like | do ...)       = [2gram] 0.180077 [ -0.744543 ]
        p( <unk> | like ...)    = [1gram] 0.0319473 [ -1.49557 ]
        p( </s> | <unk> ...)    = [1gram] 0.206897 [ -0.684247 ]
1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -4.93074 ppl= 6.63422 ppl1= 9.68608

file sample-data/test-q5.txt: 1 sentences, 5 words, 0 OOVs
0 zeroprobs, logprob= -4.93074 ppl= 6.63422 ppl1= 9.68608
```

Brown Corpus

The test dataset does not contain punctuation and start of sentences do not begin with capital letters. Hence, I used the above unix command for removing all punctuation and the capital letters in the start of the sentences in the training data set – “brown-train.txt”.

Preprocessing

In [ ]:

```
#preprocessing
!cat browndata/brown-train.txt | tr '[:upper:]' '[:lower:]' | sed -e "s/[[:punct:]]\+//g" >
```

In [ ]:

```
#creating count file
!./ngram-count -text browndata/brown-train.txt -order 2 -write brown-train.count
```

In [ ]:

```
# language modelling (without -preprocessing)
!./ngram-count -text browndata/brown-train.txt -order 2 -write brown-train.count -unk -lm b
```

In [ ]:

```
# language modelling (with -preprocessing)
!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm
```

In [ ]:

```
# language modelling testing (without -preprocessing)
!./ngram -lm brown-train.lm -ppl browndata/brown-test.txt
```

```
file browndata/brown-test.txt: 14334 sentences, 305056 words, 18639 OOVs
39856 zeroprobs, logprob= -990704 ppl= 6270.9 ppl1= 10425.3
```

In [ ]:

```
# language modelling testing (with -preprocessing)
!./ngram -lm brown-train-pre.lm -ppl browndata/brown-test.txt
```

```
file browndata/brown-test.txt: 14334 sentences, 305056 words, 87046 OOVs
26120 zeroprobs, logprob= -754354 ppl= 4549.2 ppl1= 8534.52
```

order 1

In [ ]:

```
# language modelling order 1 (with -preprocessing)
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm
```

In [ ]:

```
# language modelling testing (with -preprocessing)
!./ngram -lm brown-train-pre-o1.lm -ppl browndata/brown-test.txt
```

```
file browndata/brown-test.txt: 14334 sentences, 305056 words, 87046 OOVs
0 zeroprobs, logprob= -687741 ppl= 912.033 ppl1= 1427.67
```

order 2

In [ ]:

```
# language modelling order 2 (with -preprocessing)
!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm
```

In [ ]:

```
# language modelling testing (with -preprocessing)
!./ngram -lm brown-train-pre-o2.lm -ppl browndata/brown-test.txt
```

```
file browndata/brown-test.txt: 14334 sentences, 305056 words, 87046 OOVs
26120 zeroprobs, logprob= -754354 ppl= 4549.2 ppl1= 8534.52
```

order 3

In [ ]:

```
# language modelling order 3 (with -preprocessing)
#!./ngram-count -text browndata/brown-train1.txt -order 3 -write brown-train.count -unk -lm
```

In [ ]:

```
# language modelling testing (with -preprocessing)
!./ngram -lm brown-train-pre-o3.lm -ppl browndata/brown-test.txt
```

```
file browndata/brown-test.txt: 14334 sentences, 305056 words, 87046 OOVs
26120 zeroprobs, logprob= -752526 ppl= 4457.31 ppl1= 8349.39
```

order 4

In [ ]:

```
# language modelling order 4 (with -preprocessing)
#!./ngram-count -text browndata/brown-train1.txt -order 4 -write brown-train.count -unk -lm
```

In [ ]:

```
# language modelling testing (with -preprocessing)
!./ngram -lm brown-train-pre-o4.lm -ppl browndata/brown-test.txt
```

```
file browndata/brown-test.txt: 14334 sentences, 305056 words, 87046 OOVs
26120 zeroprobs, logprob= -752526 ppl= 4457.31 ppl1= 8349.39
```

order 5

In [ ]:

```
# language modelling order 5 (with -preprocessing)
#!./ngram-count -text browndata/brown-train1.txt -order 5 -write brown-train.count -unk -lm
```

In [ ]:

```
# language modelling testing (with -preprocessing)
!./ngram -lm brown-train-pre-o5.lm -ppl browndata/brown-test.txt
```

```
file browndata/brown-test.txt: 14334 sentences, 305056 words, 87046 OOVs
26120 zeroprobs, logprob= -752526 ppl= 4457.31 ppl1= 8349.39
```

we have

1. brown training set
2. brown testing set
3. brown devlopment set

Find out best lamba [0,1] (also known as interpolation weight) and -addsmoothparameters for which you get less perplexity on your development set.This process is also known as hyperparameter tuning.

In [ ]:

```
#Lambda = 0.1

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.1 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -287469 ppl= 713.237 ppl1= 1060.27
```

In [ ]:

```
#Lambda = 0.2

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.2 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -282810 ppl= 641.194 ppl1= 947.068
```

In [ ]:

```
#Lambda = 0.3

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.3 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -280085 ppl= 602.486 ppl1= 886.558
```

In [ ]:

```
#Lambda = 0.4

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.4 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -278522 ppl= 581.34 ppl1= 853.599
```

In [ ]:

```
#Lambda = 0.5

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -277872 ppl= 572.765 ppl1= 840.255
```

In [ ]:

```
#Lambda = 0.6

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.6 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -278108 ppl= 575.865 ppl1= 845.077
```

In [ ]:

```
#Lambda = 0.7

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.7 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -279404 ppl= 593.174 ppl1= 872.035
```

In [ ]:

```
#Lambda = 0.8

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.8 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -282313 ppl= 633.951 ppl1= 935.729
```

In [ ]:

```
#Lambda = 0.9

!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.9 -ppl browndata/brown-dev.txt
```

```
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 31813 OOVs
0 zeroprobs, logprob= -288764 ppl= 734.667 ppl1= 1094.08
```

lambda = 0.5 has minimum perplexity

now varying smoothing parameters

In [ ]:

```
#addsmooth 1

#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -470970 ppl= 3570.67 ppl1= 5168.55
```

In [ ]:

```
#addsmooth 2

#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -474921 ppl= 3824.31 ppl1= 5552.89
```

In [ ]:

```
#addsmooth 3

#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -478061 ppl= 4038.69 ppl1= 5878.65
```

In [ ]:

```
#addsmooth 4

#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -476570 ppl= 3935.42 ppl1= 5721.63
```

In [ ]:

```
#addsmooth 5

#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -475846 ppl= 3886.21 ppl1= 5646.87
```

optimal parameter is addsmooth 1 and lambda = 0.5

In [ ]:

```
#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -470970 ppl= 3570.67 ppl1= 5168.55
```

In [ ]:

```
#Testing on test set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-test.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-test.txt: 14334 sentences, 305056 words, 0 OOVs
0 zeroprobs, logprob= -1.15701e+06 ppl= 4193.46 ppl1= 6205.69
```

Good tunning smoothing

In [ ]:

```
#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
warning: discount coeff 1 is out of range: 0
warning: discount coeff 1 is out of range: 0
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -450373 ppl= 2496.74 ppl1= 3556.04
```

In [ ]:

```
#Testing on test set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-test.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-test.txt: 14334 sentences, 305056 words, 0 OOVs
0 zeroprobs, logprob= -1.11175e+06 ppl= 3025.88 ppl1= 4409.71
```

Kneser-Ney smoothing.

In [224]:

```
#Training bigram model on the train set
!./ngram-count -text browndata/brown-train1.txt -order 1 -write brown-train.count -unk -lm

!./ngram-count -text browndata/brown-train1.txt -order 2 -write brown-train.count -unk -lm

#Testing on dev set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-dev.txt
```

```
warning: discount coeff 1 is out of range: 0
warning: discount coeff 1 is out of range: 0
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-dev.txt: 5734 sentences, 126831 words, 0 OOVs
0 zeroprobs, logprob= -450373 ppl= 2496.74 ppl1= 3556.04
```

In [225]:

```
#Testing on test set
!./ngram -lm /content/brown-train-pre-o2.lm -mix-lm /content/brown-train-pre-o1.lm \
-lambda 0.5 -ppl browndata/brown-test.txt
```

```
/content/brown-train-pre-o2.lm: line 939: warning: non-zero probability for
<unk> in closed-vocabulary LM
/content/brown-train-pre-o1.lm: line 938: warning: non-zero probability for
<unk> in closed-vocabulary LM
file browndata/brown-test.txt: 14334 sentences, 305056 words, 0 OOVs
0 zeroprobs, logprob= -1.11175e+06 ppl= 3025.88 ppl1= 4409.71
```

In [ ]: