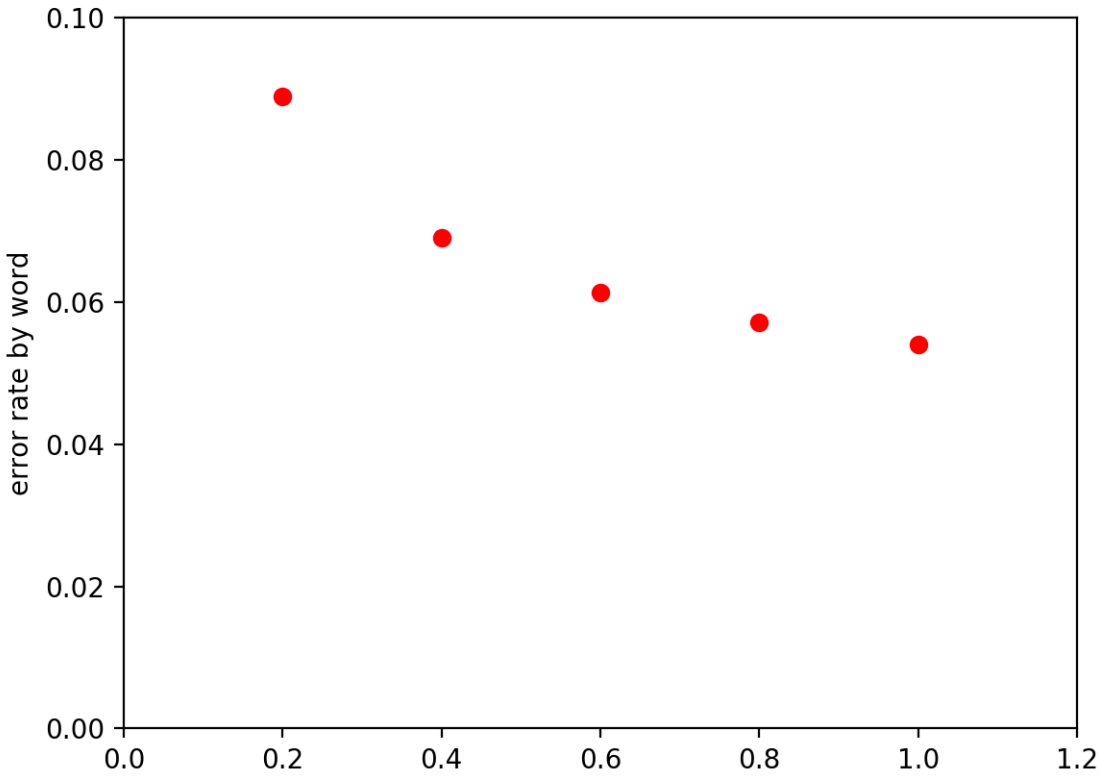
Dept. of Computer Science, University of California, Davis
ECS189G Weiran Guo(912916431) February 13, 2018

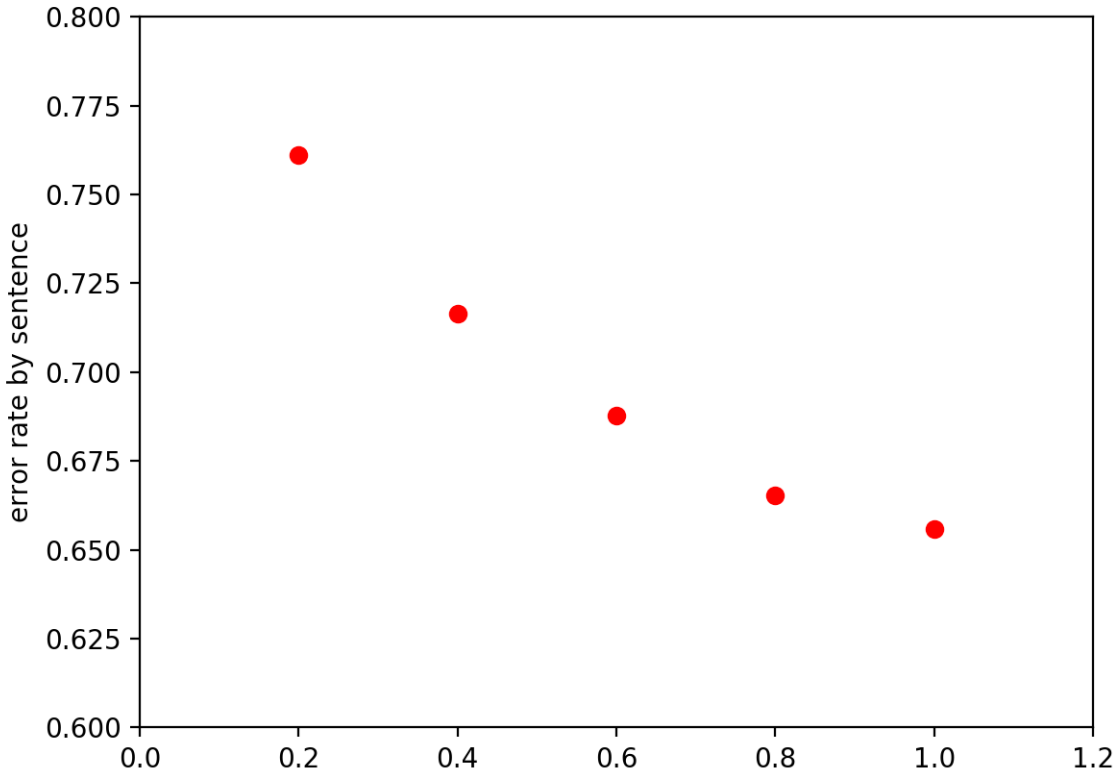
HW2 Report

1 Task 1

the error rate by word relative to the corpus



x is the multiple of the corpus size(39832)



x is the multiple of the corpus size(39832)

We can see that, as the training data increases, the learning rate becomes smaller, which means the training efficiency decreases. So it takes larger dataset to improve the same amount of accuracy as data grows.

2 Task 2

At first, I used two methods for training.
One is smoothed Bigram, another is Trigram without smoothing. As i found that Trigram takes a longer time and higher error rates than smoothed bigram. So it's better to use smoothed Bigram for this problem.

The smoothing method: each time meet a unknown token, instead of setting it to OOV WORD, I let it be the original word, so there are more informative transition and emission cases, at the same time, i add all the emission tags by 1, i.e. add OOV WORD inside each tag of emission, hence smooth the model.

for the development data ptb.22,
The Original HMM model given
error rate by word: 0.0540917815389984 (2170 errors out of 40117)
error rate by sentence: 0.655882352941176 (1115 errors out of 1700)

My model gives
error rate by word: 0.0489069471795 (1962 errors out of 40117)
error rate by sentence: 0.630588235294118 (1072 errors out of 1700)

After running my tagger on ptb.23.txt, the my.out file is included in the directory.

3 Task 3

for the baseline:
Japanese:
error rate by word: 0.0628611451584661 (359 errors out of 5711)
error rate by sentence: 0.136812411847673 (97 errors out of 709)
Bulgarian:
error rate by word: 0.115942028985507 (688 errors out of 5934)
error rate by sentence: 0.751256281407035 (299 errors out of 398)

My Model:
Japanese:
error rate by word: 0.0618105410611101 (353 errors out of 5711)
error rate by sentence: 0.133991537376587 (95 errors out of 709)
Bulgarian:
error rate by word: 0.0930232558139535 (552 errors out of 5934)
error rate by sentence: 0.648241206030151 (258 errors out of 398)

Sentence Error table:

Error rate	English	Bulgarian	Japanese
Baseline	0.655882352941176	0.751256281407035	0.136812411847673
Mine	0.630588235294118	0.648241206030151	0.133991537376587

As error rates shown, there are big improvement on the Bulgarian, small improvement on English(task 2 shown), and tiny improvement on Japanese.

Data Analyzing:

```
[guowr97@pc33 hw2]$ wc -l *.hmm
  34722 btb.hmm
   5854 jv.hmm
  53125 my.hmm
```

It shows that there are fewer bigram POS patterns in Japanese and huge bigram patterns in English and Bulgarian.

There is one possible reasons for the big improvement on the Bulgarian:
1. the baseline error rate is high, hence there are much space for improvement.
There are few possible reasons for big error rate of Bulgarian:
1. the bigram model is not suit for the Bulgarian(use N-gram instead)
2. the training dataset is too small
one possible reasons for the small improvement and big error rate on the English
1. the bigram model is not suit for English(N-gram instead)
Possible reasons for low error rate and tiny improvement on Japanese:
1.Bigram model is suit for Japanese
2.Since the error rate is low, there is limited space for improvement