

Hw2: Language Modeling of Different Corpora by Domain and Timeframe

Sam Showalter

University of California, Irvine (showalte)

Kaggle: Sam Showalter

showalte@uci.edu

	brown	reuters	gutenberg
uni.	1514 / 1758	1467 / 1577	981 / 1036
add-k	128 / 1822	72 / 378	124 / 650
b-off	75 / 559	51 / 180	92 / 285

Table 1: Hyperparameter tuning results. Dev-set perplexity across all three corpora with different smoothing. add-k and backoff (b-off) running optimal $k=0.01$ and $n=2$ for ngram models

Abstract

SOMETHING (Chen and Goodman, 1999)

1 Related Work: Language Model Implementations

SOMETHING

2 Experimental Setup

SOMETHING

3 Hyperparameter Tuning

3.1 In-Domain Text Analysis: Empirical

3.2 Out-of-Domain Text Analysis: Empirical

	brown	reuters	gutenberg
brown	558.59	2118.33	840.811
reuters	1500.02	180.312	2159.68
gutenberg	1098.51	4240.6	285.342

Table 2: Test set perplexity for each language model with stupid backoff smoothing applied. Each model is applied to every corpus.

3.3 Qualitative Analysis: Language and Generation Scoring

Prefix: It is

Brown:

sent.	brown	reuters	gutenberg
0	6836 / 5268	490 / 1571	4012 / 6455
1	1901 / 1997	1765 / 2670	519 / 1325
2	621 / 1081	5685 / 2628	2094 / 1572
3	1011 / 687	1353 / 1334	1884 / 1473
4 ^g	1270 / 1663	2296 / 2656	124 / 1271
5 ^r	110 / 335	245 / 882	111 / 361
6 ^r	1498 / 2288	846 / 1430	3863 / 4466

Table 3: Perplexity scoring of sentences (indexed in appendix A) with best backoff smoothing language model. Columns represent the language model that created the perplexity score. Superscores represent generating model, if not human.

Reuters:

Gutenberg:

4 Conclusion

SOMETHING

5 Statement of Collaboration

SOMETHING

References

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

A Sentence Reference

Below is a reference for table with generated sentences. As noted above, the sentences indexed in 3 with an additional superscript were generated by the reuters model while *g* references gutenberg.

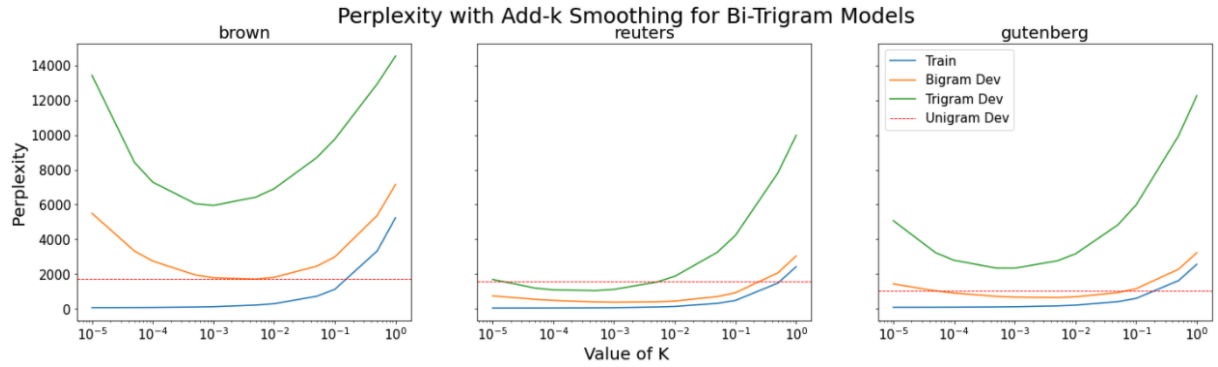


Figure 1: Train and dev-set perplexity for bi- and tri-gram language models with vary k in add-k smoothing. A $k=1$ represents Laplacian smoothing while dropping k represents weakening the Dirichlet prior based on the data. Referenced is dev-set perplexity for unigram models.

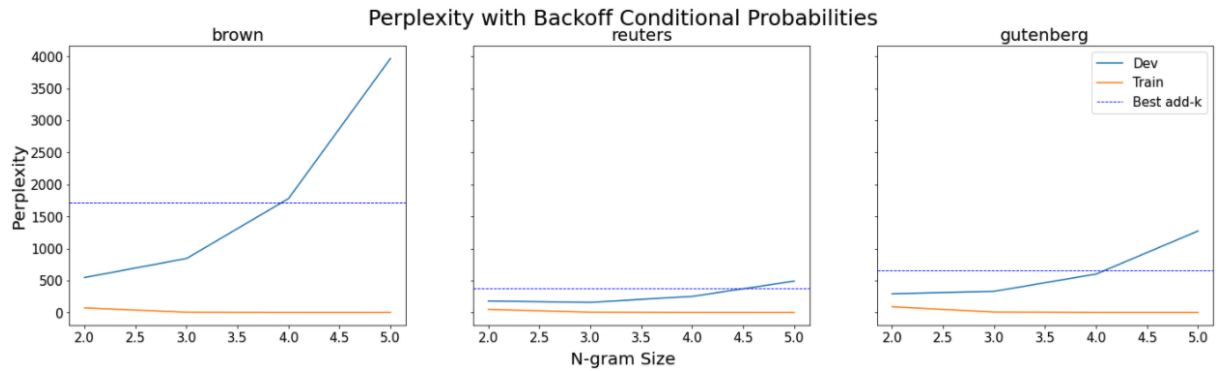


Figure 2: Dev-set perplexity across corpora with stupid backoff smoothing implemented with $\lambda = 0.4$ and varying number of context sizes (n-gram length). Referenced is current best performance from add-k smoothing (dotted).

index	sentence
0	This third quarter fiscal forecast for DOW Chemical is bearish according to financial analysts.
1	Who hath such scuples as to remain untrodden by the perils of temptation.
2	Hey did hear about that fight the football team got in to after practice?
3	Walked alone did she, on to tomorrow.
4 ^g	It was like fire sent them and new psychological influences set in stone brought the Holy Ghost
5 ^r	It was like the workforce body
6 ^r	It was like a loss for Bahrain Oman and current residents are packing said company spokesman