Mark Min & Remy Zachwieja
6/9/21
CSE/LING 472
M4 Write-up

Introduction:

An n-gram language model is a model that assigns probabilities to sentences. It does this through n-gram probabilities which use conditional probabilities based on the previous n-1 words in order to determine the likelihood of a word occurring. The probabilities are normalized in order to not heavily favor words that repeat themselves a lot and words that don't appear at all or an insignificant number of times. We implemented two smoothing algorithms. For unigram and bigram, we use Laplace smoothing which means we use an Add One Estimation instead of a Maximum Likelihood Estimation to help deal with unseen n-gram probabilities in our dataset. For trigram we implemented stupid backoff which dealt with unseen trigrams by checking for a bigram possibility looking only for n given n-1. Then if we couldn't find a bigram possibility, we would use our unigram probability. Every time we would go down an n-gram we would multiply our probability by a suggested alpha value 0.4 in order to offset the fact that we hadn't seen that specific trigram or bigram. Perplexity is a way to determine how effective a language model is when used on a test set. Perplexity is calculated as the inverse of the probability of a sentence occurring so when run on a test set, a lower perplexity means more likely sentences, which means our language model is more effective.

Data Statement:

Curation Rationale: Our data set consisted of sentences gathered from Jane Austen's works. These works were selected for us.

Language Variety: English as spoken in late 1700's and early 1800's in England.

Speaker Demographic: Jane Austen was a relatively upper-class women, brought up in late 1700's England.

Annotator Demographic: This language model uses more of Jane Austen's works as our test text so there isn't much annotator influence aside from the initial choice of using Austen's works.

Speech Situation: Late 1700's England, Written, Scripted, asynchronous interaction and intended for other upper class Christian English women.

Text Characteristics: Austen is known for her critique of British society with a focus of love in her works so most of her novels have these common themes.

Recording Quality: Written, so recorded perfectly and likely edited to her preferences.

Results:

Perplexities for our different Models trained on austen_train set

	Unigram (Laplace Smoothing)	**Bigram (Laplace Smoothing)**	**Trigram (Stupid Backoff)**
Train	432.424	405.588	7.174
Dev	416.246	473	108.783
Test	416.807	478.076	109.558

Our perplexities for unigram and bigram are relatively high which is to be expected, but bigram perplexity being higher than unigram for dev and test sets. We assume that this is because of a relatively high number of unseen bigrams appearing when evaluating the dev and test set. We are using Laplace smoothing, so whenever an unseen bigram it's assigned a fairly low probability. The probabilities of these unseen bigrams may be what results in an overall probability that is lower (higher perplexity) for our bigram model. This just goes to show that Laplace smoothing is not an ideal smoothing method for Language Models like this. Our trigram probability looks reasonable however as it is significantly lower likely due to a combination of Stupid Backoff being used instead of Laplace and that trigrams should be a better predictor of sentences than bigrams and unigrams.

References:

Bender, E., & Friedman, B. (2018). “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science” [Review of “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science”]. Transactions of the Association for Computational Linguistics, 6, 587--604. https://doi.org/10.1162/tacl_a_00041, https://www.aclweb.org/anthology/D07-1090.pdf

Brants, T., Popat, A., Xu, P., Och, F., & Dean, J. (2007). Large Language Models in Machine Translation (pp. 858–867). Association for Computational Linguistics. https://www.aclweb.org/anthology/D07-1090.pdf

Daniel, J., & Martin, J. (2020). Speech and Language Processing - Chapter 3 N-gram Language Models. https://web.stanford.edu/~jurafsky/slp3/3.pdf