CPSC 436N - Assignment 1

A1 27: Nima Kondori, Victoria Wu, Hooman Vaseli

Due Date: Oct 12th, 2023

Count-based Model

Q1

For the bigram model, what is **P**(agency| the) when the model is built with **data_mini**? What is it when the model is built with data?

Answer:
Using data: 0.0019393651547534236
Using **data_mini**: 0.0011614401858304297

Q2

Compare and report the performance (perplexity on the test set) of the bigram and the trigram models, and provide a possible explanation for what you find.

Answer:
The perplexity of the test set using the bigram model was **88.32** while that of the trigram model was **215.18**. The perplexity of the trigram model is higher than that of the bigram model. While we expect the trigram model to have lower perplexity because it can take more context into account, there are a few reasons why the perplexity may be higher. One reason could be the sparsity/low size of the training set. Because bigrams only look at two words at a time, they require fewer words to estimate the probability accurately. However, trigrams have a larger window and thus need more training data to get good probability estimations. Another reason could be that the sentences in the corpus have some ambiguity or are shorter in length, so introducing additional context does not actually help the model calculate the probabilities. Instead, relying on a single previous word allows the model to perform better.

Neural Model

Q3

Why is the output dimension of the final layer vocab_size?

Answer:
The neural network predicts the probability of the next token. Since it would be any of the tokens in our vocabulary, our output dimension must capture all the possible options. This results in a vector of vocab_size, where each element in the vector represents a word in the vocab. This is then used in training using the cross-entropy loss, which uses the normalized probabilities of the vocabulary against the ground-truth word token to improve the model.

Comparing the two language models

Q4

Now compare the obtained perplexity for your best neural model with that of the best count-based model. Which one achieves lower perplexity? Can you provide an explanation for this finding?

Answer:
Comparing the perplexity of the bigram (88.32), best count-based model, against the best neural model (18.30), we see that the neural model is superior.
This can be explained by several reasons: (1) The neural network uses an embedding layer that learns to embed the tokens with similar semantics and meanings close to each other which will make the use of them in future layers better. (2) But most importantly, the neural network approach is different than count based because it learns a set of weights to estimate the posterior probability of the next word given a n-gram context, rather than being dependant on counting the n-grams in the corpus which may be hindered by the size of the corpus.
This allows the neural model to generalize unseen data better, as it does not need to see a previous instance of an n-gram to estimate the probability accurately. Because the test set may contain unseen n-grams, the neural model performs better on it.

Q5

Compare the text generated by the best count-based model and the neural model. Which one is better and in what aspects?

Answer:
When we compare the count-based model and neural model, the sentences generated by both models are not very accurate or meaningful. We see that for k=1 the generated sentences are actually the same. This can be attributed to the fact that **<unk>** token is the best one predicted by both models so it is always selected. When k=1000, both models generated sentences that do not make sense. The count-based model suggested a more diverse variety of tokens. However, this means it often suggested tokens that did not make sense in the context of the sentence, like "N" or "$" or numbers. Although the sentences suggested by the neural model did not necessarily make sense either, they were more consistent in terms of vocabulary and style. The tokens seemed less random, and the sentences maintained a more consistent tone.

Q6

Please indicate whether you used generative AI tools like ChatGPT for this assignment and how. This information is for data collection purposes and will not be used against you in any way.

Answer:
No we didn't use ChatGPT or similar AI tools (e.g. GitHub copilot, etc).