Sharvita Paithankar
Report for NLP Assignment 1

1. What do you expect the difference between the Brown bigram and trigram models to look like? Which model will provide you with more coherent text? How will the perplexity of each compare? You should test your predictor and perplexity function using the brown_bigrams and brown_trigrams to confirm your expectations. For perplexity, an average over 2-5 sentences from the Brown corpus should be fine, but make sure you use the same sentences both times. If something you did not expect occurs, explain what happened and why you believe it happened.

Usually the trigram model generates more coherent text than the bigram model. This is because the trigram adds an additional level of context. It uses two previous words to predict the next word rather than just using one. Using the idea of adding to the context should result in more natural and structured sentence.

When testing my predictor function, we observed the following generated sequences:

- **Bigram Generated Sentence:** "The `` in the page is not work in the page."
- **Trigram Generated Sentence:** "The first of the American League 's 1961 expansion to the public."

The trigram model had more meaningful sentences, but the bigram model struggled with coherence,

It ended up looping or generating sentence that did not make sense grammatically.

For perplexity, we calculated scores over selected sentences:

- **Brown Bigram Perplexity:** 104.857
- **Brown Trigram Perplexity:** 4.323

Perplexity is lower for the trigram model, meaning that it assigns higher probabilities to observed sequences and is thus a better fit for the data. The bigram model, with its higher perplexity, struggles more to predict words accurately, confirming our expectations. The trigram model benefits from additional context, making it more confident in its predictions.

However, in some cases, if the corpus is too small, trigrams may overfit to specific word sequences, reducing their generalization ability. This could lead to unexpected results, such as a model being unable to generate diverse outputs due to insufficient data supporting three-word sequences.

2. When testing our bigram models on the Reuters data, do you think a model trained on Brown or Webtext will perform best? Pick any 25 sentences from the Reuters corpus and calculate the average perplexity using each of your bigram datasets. Compare the results of each and provide explanation as to why you believe that one performed better than the other.

To figure out whether a model trained on Brown or Webtext would perform better on Reuters, I evaluated the average perplexity on 25 randomly chosen sentences:

- **Brown BiGram Perplexity on Reuters:** 88.34
- **WebText BiGram Perplexity on Reuters:** 108.09

The Brown bigram model performed better and had lower perplexity than the WebText bigram model when tested on Reuters data. This means that the Brown corpus shares more linguistic similarities with Reuters than WebText does. Since the Brown corpus consists of well-structured, formal English texts across multiple domains, it provides better generalization for structured content like Reuters news articles.

WebText has more informal, conversational, and varied content, which does not not align well with the structured nature of Reuters articles. This explains why the WebText model has a higher perplexity.

3. When predicting the next word in a sentence, what do you believe would happen if we increased the number of sentences in our training data?

Increasing the number of sentences in our training data will likely lead to improvements in model performance. It will have more robust predictions and more training data provides the model with a larger variety of word combinations. It reduces the chance of encountering unseen word pairs or triplets. A large dataset allows for better probability estimations and leads to a model with lower perplexity scores. The model also becomes more versatile in predicting unseen text, reduces overfitting to specific phrases seen in a smaller corpus.

But there are also downsides. If the dataset is too large and contains noisy or irrelevant text, it may show errors and make the model less efficient. Additionally, training time increases significantly as more data needs to be processed and stored.

The trigram model outperforms the bigram model in coherence and perplexity due to the additional context it considers. With Reuters data, the Brown bigram model performed better than the WebText model, probably because its more formal and structured nature. Increasing the size of our training corpus should generally improve performance, provided the data remains relevant and representative of the target domain.