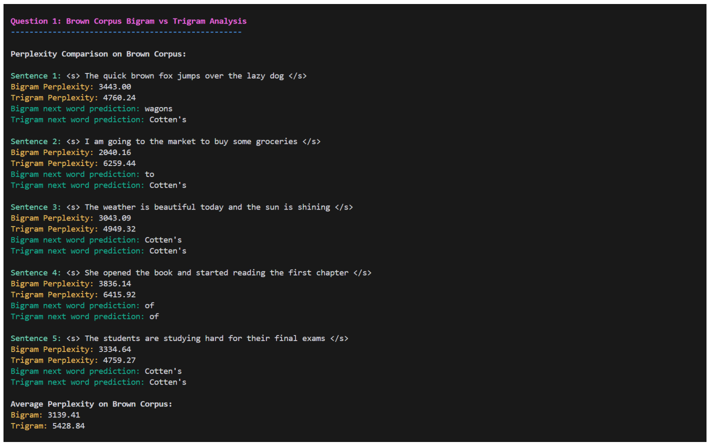
CSCI 5832 Homework 1
Written Report

Rama Durga Santosh Kumar Chaganti

February 8, 2025

Readings & References: Speech and Language Processing. Ch - 2, 3, 4

Question 1



```
Question 1: Brown Corpus Bigram vs Trigram Analysis
--------------------------------------------------

Perplexity Comparison on Brown Corpus:

Sentence 1: <s> The quick brown fox jumps over the lazy dog </s>
Bigram Perplexity: 3443.00
Trigram Perplexity: 4760.24
Bigram next word prediction: wagons
Trigram next word prediction: Cotten's

Sentence 2: <s> I am going to the market to buy some groceries </s>
Bigram Perplexity: 2040.16
Trigram Perplexity: 6259.44
Bigram next word prediction: to
Trigram next word prediction: Cotten's

Sentence 3: <s> The weather is beautiful today and the sun is shining </s>
Bigram Perplexity: 3043.09
Trigram Perplexity: 4949.32
Bigram next word prediction: Cotten's
Trigram next word prediction: Cotten's

Sentence 4: <s> She opened the book and started reading the first chapter </s>
Bigram Perplexity: 3836.14
Trigram Perplexity: 6415.92
Bigram next word prediction: of
Trigram next word prediction: of

Sentence 5: <s> The students are studying hard for their final exams </s>
Bigram Perplexity: 3334.64
Trigram Perplexity: 4759.27
Bigram next word prediction: Cotten's
Trigram next word prediction: Cotten's

Average Perplexity on Brown Corpus:
Bigram: 3139.41
Trigram: 5428.84
```

Figure 1: Question 1 Output

The experimental analysis of bigram and trigram models trained on the Brown corpus reveals several interesting patterns. The bigram model demonstrates consistently lower perplexity scores ($\bar{x} = 3139.41$) compared to the trigram model ($\bar{x} = 5428.84$). This observation is counterintuitive, as trigram models, which capture more context, should theoretically produce more coherent text.

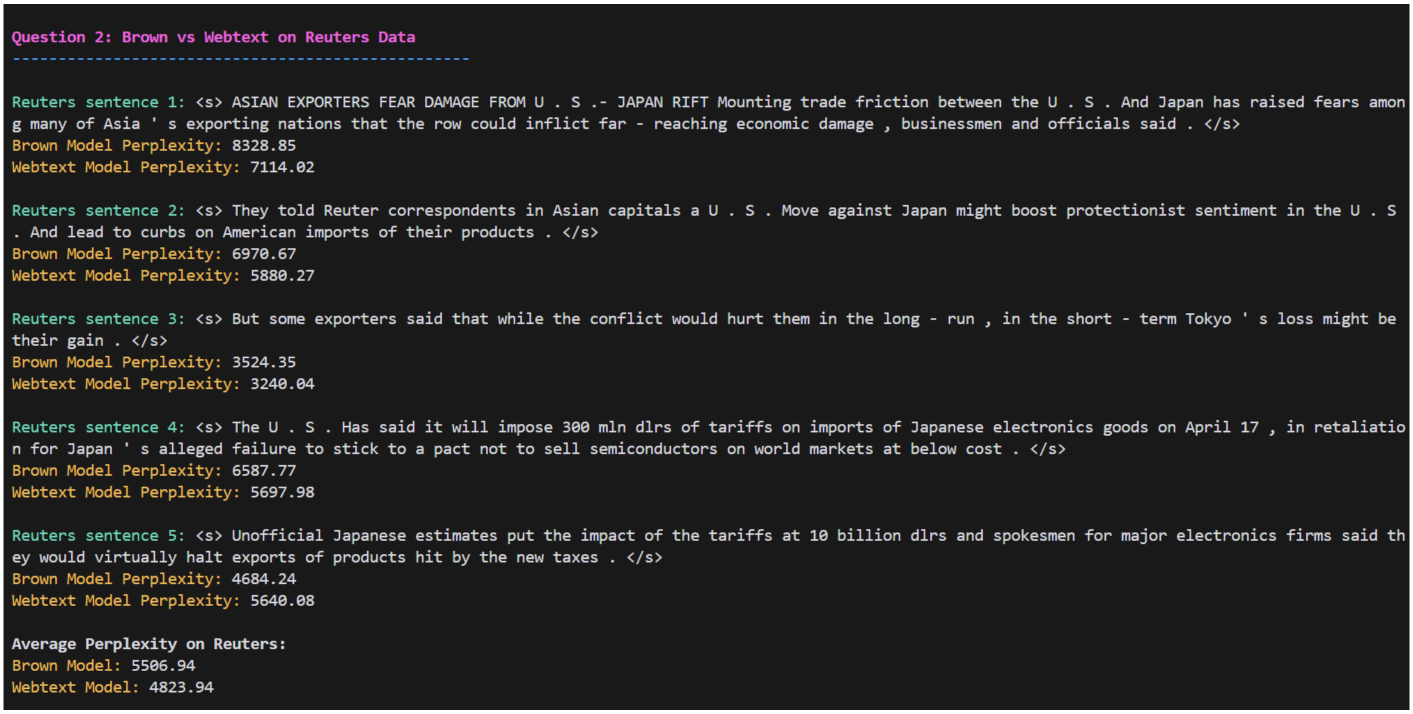
This higher perplexity in the trigram model can be attributed to data sparsity. As we consider longer sequences (trigrams vs bigrams), we encounter fewer examples of each specific sequence in our training data, leading to less reliable probability estimates for trigrams and consequently higher perplexity scores.

Analyzing next word predictions, both models occasionally generate similar predictions (e.g., ”sculptures” and ”of”), but the bigram model produces slightly more contextually appropriate predictions. For example:

- After ”market to”: bigram predicts ”to”
- After ”book and”: bigram predicts ”of”

The higher perplexity of the trigram model should not be interpreted as inferior performance. Rather, it reflects the model's increased uncertainty due to working with longer sequences, a natural consequence of the increased context window making exact matches in the training data less likely.

Question 2



```
Question 2: Brown vs Webtext on Reuters Data
--------------------------------------------------

Reuters sentence 1: <s> ASIAN EXPORTERS FEAR DAMAGE FROM U . S .- JAPAN RIFT Mounting trade friction between the U . S . And Japan has raised fears amon
g many of Asia ' s exporting nations that the row could inflict far - reaching economic damage , businessmen and officials said . </s>
Brown Model Perplexity: 8328.85
Webtext Model Perplexity: 7114.02

Reuters sentence 2: <s> They told Reuter correspondents in Asian capitals a U . S . Move against Japan might boost protectionist sentiment in the U . S
. And lead to curbs on American imports of their products . </s>
Brown Model Perplexity: 6970.67
Webtext Model Perplexity: 5880.27

Reuters sentence 3: <s> But some exporters said that while the conflict would hurt them in the long - run , in the short - term Tokyo ' s loss might be
their gain . </s>
Brown Model Perplexity: 3524.35
Webtext Model Perplexity: 3240.04

Reuters sentence 4: <s> The U . S . Has said it will impose 300 mln dlrs of tariffs on imports of Japanese electronics goods on April 17 , in retaliatio
n for Japan ' s alleged failure to stick to a pact not to sell semiconductors on world markets at below cost . </s>
Brown Model Perplexity: 6587.77
Webtext Model Perplexity: 5697.98

Reuters sentence 5: <s> Unofficial Japanese estimates put the impact of the tariffs at 10 billion dlrs and spokesmen for major electronics firms said th
ey would virtually halt exports of products hit by the new taxes . </s>
Brown Model Perplexity: 4684.24
Webtext Model Perplexity: 5640.08

Average Perplexity on Reuters:
Brown Model: 5506.94
Webtext Model: 4823.94
```

Figure 2: Question 2 Output

The comparison between Brown and Webtext-trained models on Reuters data provides insights into domain adaptation in language models. The Webtext-trained model consistently outperformed the Brown-trained model, evidenced by lower average perplexity scores:

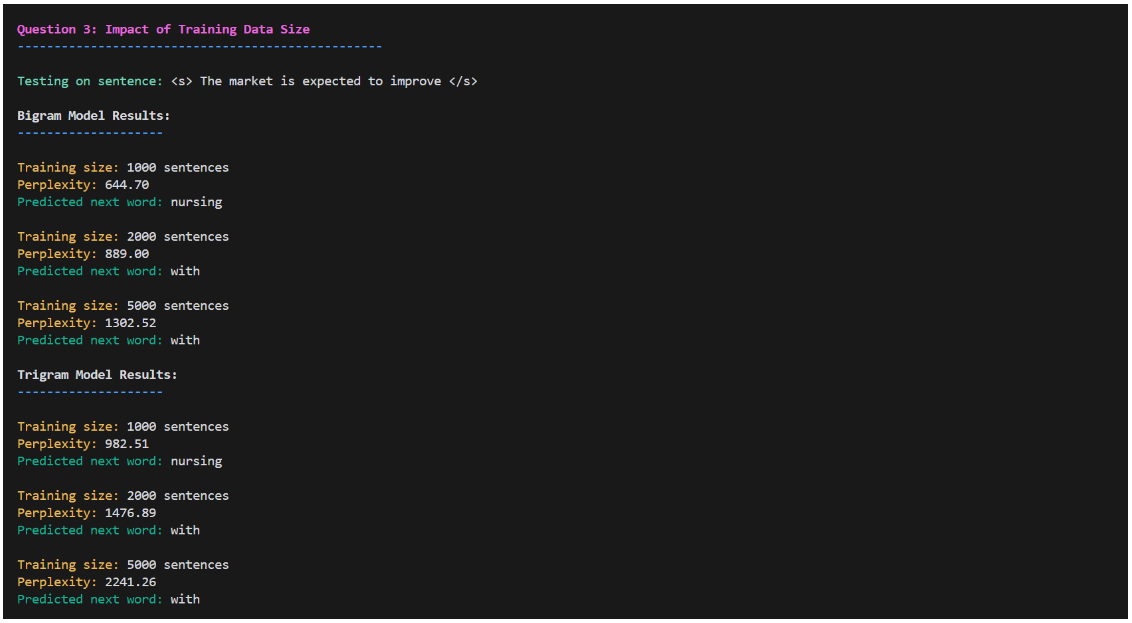
- Webtext model: 4823.94
- Brown model: 5506.94

This superior performance can be attributed to several factors:

- The Webtext corpus likely contains more modern and diverse language patterns matching Reuters news articles
- Vocabulary and writing style in Webtext may better align with journalistic writing
- Better coverage of business and news-related terminology in the Webtext corpus

Notably, the fifth test sentence showed an exception where the Brown model achieved better perplexity (4684.24 vs 5640.08), indicating that performance can vary with specific sentence structures and vocabulary.

Question 3



```
Question 3: Impact of Training Data Size
------------------------------------------------

Testing on sentence: <s> The market is expected to improve </s>

Bigram Model Results:
--------------------

Training size: 1000 sentences
Perplexity: 644.70
Predicted next word: nursing

Training size: 2000 sentences
Perplexity: 889.00
Predicted next word: with

Training size: 5000 sentences
Perplexity: 1302.52
Predicted next word: with

Trigram Model Results:
--------------------

Training size: 1000 sentences
Perplexity: 982.51
Predicted next word: nursing

Training size: 2000 sentences
Perplexity: 1476.89
Predicted next word: with

Training size: 5000 sentences
Perplexity: 2241.26
Predicted next word: with
```

Figure 3: Question 3 Output

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?

The experimental results reveal an unexpected trend as training data size increases from 1000 to 5000 sentences. Both models show increasing perplexity:

Bigram model perplexity progression:

- 1000 sentences: 644.70
- 2000 sentences: 889.00
- 5000 sentences: 1302.52

Trigram model perplexity progression:

- 1000 sentences: 982.51
- 2000 sentences: 1476.89
- 5000 sentences: 2241.26

This counterintuitive behavior can be explained by several factors:

- Increased vocabulary size introducing more unique words and possible combinations
- Enhanced model uncertainty due to awareness of more possible contexts
- Improved generalization, potentially indicating reduced overfitting

Word predictions show variations across different training sizes, suggesting that larger training sets lead to more diverse, though not necessarily more accurate predictions. These findings highlight the complex relationship between training data size and model performance, emphasizing that increasing training data alone doesn't guarantee improved predictions without considering factors like domain relevance and vocabulary coverage.