Using N-grams as Tokens to Create Phrases

Mochalova Elizaveta

December 8, 2024

1 Introduction

In this project, we aim to develop a model that generates coherent, grammatically accurate phrases using N-gram modeling. The main process involves constructing N-grams from text data and then using these N-grams to generate phrases. We measure the quality of generated phrases using different metrics. For additional refinement, we employ the HappyTextToText transformer model to correct and polish the generated phrases, aligning them more closely with natural language.
This approach combines the strengths of N-grams for phrase creating with the grammatical capabilities of transformers, resulting in generation of high-quality phrases.

2 Construction of N-grams

2.1 Dataset for N-grams

To start, we need to prepare N-grams so that they can be used as tokens later on to create whole phrases. This can be one using one of two approaches:

- using an existing dictionary of N-grams, such as Google Ngram
- creating a new dictionary from a large corpus of text.

We will be using the second approach since it would be easier to train our model on the topics we are interested in.
Another choice that has to be made is how many items do we want to consider for the N-grams? It is possible to use N-grams of characters or of words, in this project we will use words. We will create a dictionary of 3 words using NLTK library corpora. The main disadvantage of NLTK is that its corpora are quite limited in size and in applications. But they are already cleaned and tokenized. This saves time on preprocessing. In NLTK library there are multiple corpora inside, such as Brown (sources categorized by genre, has around 1.1 mil words), Gutenberg (selection of texts from the Project Gutenberg, around 2.5 mil words), etc. We use NLTK Gutenberg and Brown corpora in this project work to compare the metrics of the two:

```
nltk.download('gutenberg')
all_gut_words = gutenberg.words()
all_br_words = brown.words()
```

If we want to work on a large-scale NLP project that requires substantial data, we might want to switch to a larger corpus like Common Crawl (hundreds of TB, but need to preprocess and clean), OpenWebText (about 40GB), or Wikipedia Dump (about 25GB). In this case we would need to:

- download the corpus;
- preprocess removing tags, non article sections;
- tokenize sentences into words.

A different process would be implemented if we were to use an available N-gram dataset, for example Google Ngram Viewer.

2.2 Building N-grams

We want to create an N-gram language model using the data prepared in the first step. NLTK library allows us to get the N-grams generated from a sequence of items, using *nltk.grams()*.
Since creating the model of N-grams with respective frequencies a special function was implemented that allows saving of the model as pickle file, to avoid the construction of the model every time.

3 Phrase generation

To generate complete sentences we start from N-1 random consecutive words from the corpus, then we predict the next word by sampling from the probability distribution of possible next words given the previous N-1 words, in the case of trigrams N=3, bigrams N=2, and fourgrams N=4. We continue this process until a phrase is of a certain length or a stop condition is met, such as punctuation. The choice of the next word can be done using one of these approaches:

- greedy - always pick the most probable word;
- random - sample words based on probability in the corpus. This approach is used to mitigate the use of some extremely common words.

To create phrases that flow better, we added some conditions, such as no repeating words in sequence, since it is something that very rarely happens in natural language.

4 Testing and evaluation

4.1 Metrics

After generating a phrase we pass it to an LLM to evaluate its grammar, the result can be taken as is or can be used to further increase the model's performance. Some of the metrics to evaluate the phrases can be:

- perplexity - the lower the perplexity the better;
- semantic coherence - use the LLM to assign a coherence score based on how logical or meaningful the phrase is;
- classifying phrases as coherent or not - fine-tune a classification head on the LLM to check whether phrases make sense;
- BLEU (Bilingual Evaluation Understudy) calculates the precision of N-grams generated by our custom model comparing them to some reference. It is then modified by a brevity penalty to account for phrases that are shorter than the reference ones. The formula is

  $$BLEU = BP * exp(\sum pn)$$

  .
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the similarity between the machine-generated phrases and the reference ones using overlapping n-grams, word sequences that appear in both. The formula is

  $$ROUGE = \sum(Recall)$$

  .

We mostly concentrated on the last two metrics, using a library called evaluate, considering references as the corrected phrases and predictions as the original phrases produced by N-gram model.

```
bleu = evaluate.load("bleu")
rouge = evaluate.load('rouge')
```

4.2 Correction of phrases

In this project work we use HappyTextToText, which is a transformer base on T5, to check and correct the grammar of the phrase produced by the N-gram model. Here it is possible to see an example of how the model corrects a given phrase:

- phrase created by the N-gram *", or D ' ye do well : And court the fair girl ' s silver penny ."*
- phrase modified by HTTT *"And court the fair girl's silver penny."*

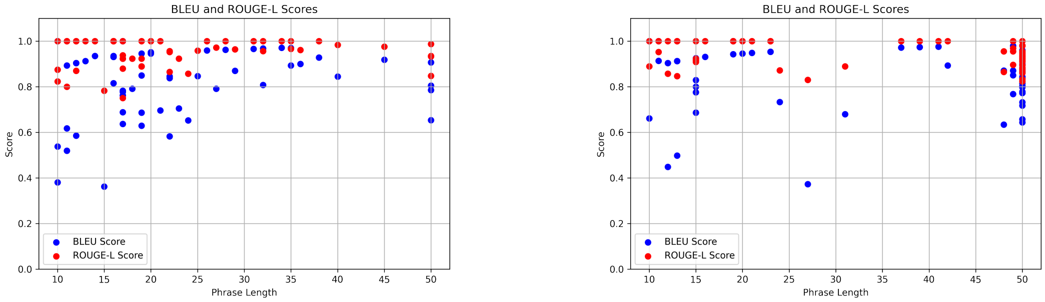
5 Results

In this project work we implemented a program that is able to create a model that composes complete sentences using N-grams, we correct these sentences with HappyTextToText transformer. We tested two corpora - Gutenberg and Brown, and different N-grams, such as bigrams, trigrams, and fourgrams to compare how well models perform. We then used the created model to produce 200 sentences of length between 10 and 50.
First, we created three models from the Gutenberg corpus, these are the obtained results for phrases' evaluations as can be seen in images 1a, 2a, and 3a. The same was done for Brown corpus with results in 1b, 2b, and 3b

Another test that was executed was to see the length of the generated phrases. The model used the limit of 50 new generated tokens after which it would stop always, but in case a punctuation such as *". ! ?"* was generated the phrase would sometimes be interrupted. As can be seen in the images 4b, 5b, 6b for Brown and 4a, 5a, 6a for Gutenberg a lot of the phrases are generated fully for 50 words, especially for Gutenberg, but for Brown the phrases' lengths are more evenly distributed. To better compare our results we can look at the table 1. Overall the values are quite close with Brown's just slightly higher in regards to BLEU, this can also be seen in the figures with all the values of the scores.

	G 2	Br 2	G 3	Br 3	G 4	Br 4
BLEU	0.811	0.829	0.842	0.883	0.838	0.857
ROUGE	0.949	0.940	0.962	0.970	0.961	0.963

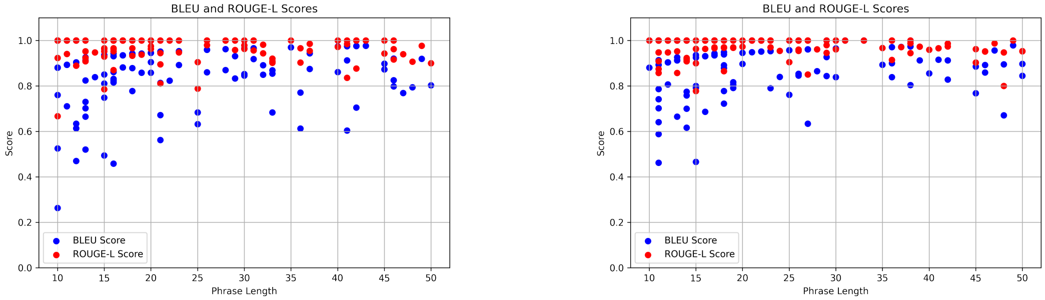
Table 1: BLEU and ROUGE metrics of different models



(a) Gutenberg corpus

(b) Brown corpus

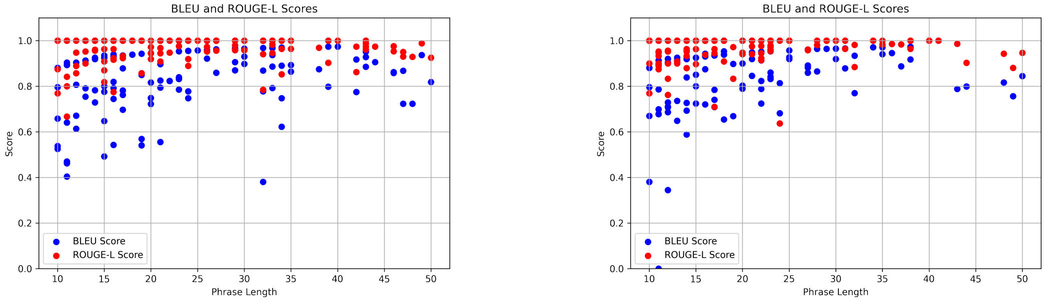
Figure 1: Phrase scores for 2-gram



(a) Gutenberg corpus

(b) Brown corpus

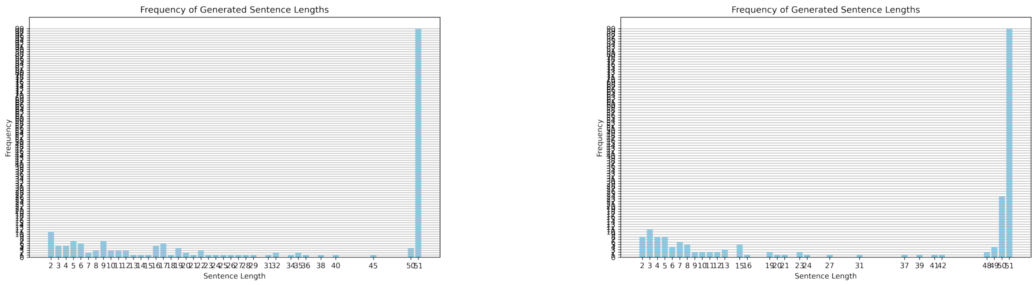
Figure 2: Phrase scores for 3-gram



(a) Gutenberg corpus

(b) Brown corpus

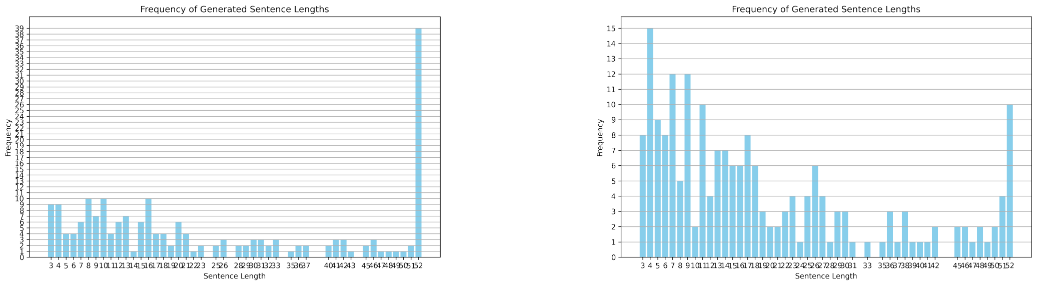
Figure 3: Phrase scores for 4-gram



(a) Gutenberg corpus

(b) Brown corpus

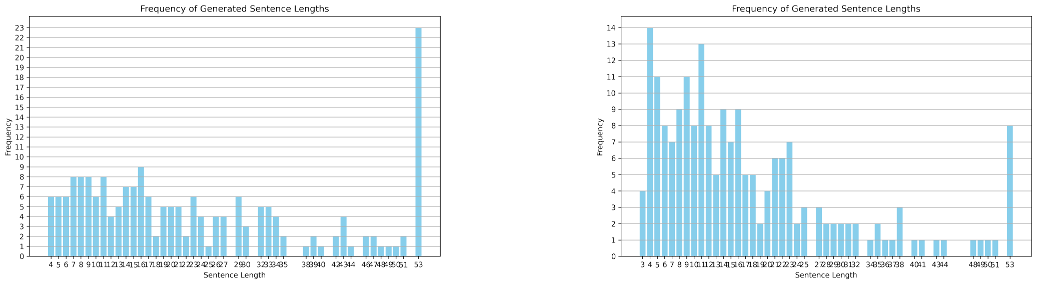
Figure 4: Length of the generated phrases for 2-gram



(a) Gutenberg corpus

(b) Brown corpus

Figure 5: Length of the generated phrases for 3-gram



(a) Gutenberg corpus

(b) Brown corpus

Figure 6: Length of the generated phrases for 4-gram

Here, we report some of the phrases generated by the N-gram models and the corrected phrases with the metrics to evaluate the produced phrases. It is important to note that HTTT does not always correct the phrase completely and it does not check how much sense the phrases make. The scores also don't account for the changes of capital letters.

Phrases using Gutenberg 2-gram model:

- precisely the similar what so as known opposite the what opposite heard the what the what.
  Exactly the similar what so as known opposite the what opposite heard the what opposite the what.
  BLEU 0.78, ROUGE 0.91

- was that disappointed buried instantly written on over quiet a inevitably the still of in at now almost strong destroyed in hurried found not now Eldad not cast sent a come speaking to often.
  Was that disappointed buried instantly written on over quiet a inevitably the still of in at now almost strong destroyed in hurried found not now Eldad not cast sent a come speaking to often.
  BLEU 0.97, ROUGE 1.0

- danger tasted, of in from, or of.
  The danger tasted, of in from.
  BLEU 0.90, ROUGE 0.93

- slow strokes, wail to, and to flapping course - But drops curling belief, and rude - rude procession to and of smile and progress of.
  Slow strokes , wail to , and to flapping course - But drops curling belief , and rude - rude procession to and of smile and progress of.
  BLEU 0.96, ROUGE 1.0

- spread their a it its forth through ensigns, in, it abroad forth his Beneath farther intensity abroad out my through under out a, upon.

Spread their a it forth through ensigns, in, it abroad forth his Beneath farther intensity abroad out my through out.
BLEU 0.80, ROUGE 0.96

Phrases using Gutenberg 3-gram model:

- of the tribe of Levi Moses gave unto the other was a disturbance to him, Elinor was hardly less, for your mother' s name was Jecholiah of Jerusalem.
  The tribe of Levi Moses gave to the other was a disturbance to him, Elinor was hardly less, for your mother's name was Jecholiah of Jerusalem.
  BLEU 0.70, ROUGE 0.95

- she had the greatest chance in life is more of the air.
  She had the greatest chance in life is more of the air.
  BLEU 0.91, ROUGE 1.0

- the Lord stood by him; and she said unto him, Micaiah, shall make supplication to the captain of the robe of the town proved all but one single star can revolve, but not equal.
  The Lord stood by him; and she said unto him, Micaiah, shall make a prayer to the captain of the robe of the town proved all but one single star can revolve, but not equal.
  BLEU 0.88, ROUGE 0.96

- flesh and the Ganges over the earth, or nought anyhow, and taught.
  The Ganges over the earth, or nought anyhow, and taught.
  BLEU 0.78, ROUGE 0.91

- got all their bodings, doubts, misgivings, fears, the family of the long and very ill of him, Art thou the chastening of the slumbering and liquid trees!
  Got all their bodings, doubts, misgivings, fears, the family of the long and very sick of him, Art thou the chastening of the slumbering and liquid trees!
  BLEU 0.89, ROUGE 0.96

Phrases using Gutenberg 4-gram model:

- You took me in," he said in a hoarse and strangled voice: " do not appear so well satisfied with my flesh ?
  You took me in," he said in a hoarse and strangled voice:" do not appear so well satisfied with my flesh?
  BLEU 1.00, ROUGE 1.00

- how slowly , but how much more natural that upon the sunniest day, if I am very small and feeble.
  How slowly , but how much more natural that on the sunniest day , if I am very small and weak.
  BLEU 0.73, ROUGE 0.89

- decent air of welcome; and Lucy, who looked at them with an overthrow, making them an ensample unto those that after should live ungodly; 2: 10 When thou comest into thy kingdom.
  And Lucy, who looked at them with an overthrow, making them an example to those that after should live ungodly; When thou comest into thy kingdom.
  BLEU 0.74, ROUGE 0.87

- herself; of being restored to Kellynch, calling it her home again, very pleas'd me, giving others the same chances and rights as myself
  She; being restored to Kellynch, calling it her home again, very pleas'd me, giving others the same chances and rights as myself.
  BLEU 0.80, ROUGE 0.96

- your assistance doe make loue, Masking the Businesse from the common decorum of a gentleman.
  Your assistance does make loue, masking the business from the common decorum of a gentleman.
  BLEU 0.52, ROUGE 0.87

Phrases using Brown 2-gram model:

- radical Bryan vapor Bryan views changes or propaganda press change to.
  Bryan vapor Bryan views changes or propaganda press changes to.
  BLEU 0.67, ROUGE 0.86
- kitchen and to and, and side, to.
  Kitchen and to and, and side, to.
  BLEU 0.88, ROUGE 1.0
- hated, the him the, her the him that name them Lublin Dolores you to the, name it him the, him that the him Dolores her the you Dolores the her it Dolores more Donald dictator the, Dolores this being them the Donald more him to
  Hated her the him that name them Lublin Dolores you to the him that Dolores her the you Dolores the her it Dolores more Donald dictator the
  BLEU 0.77, ROUGE 0.89
- Ma wasn't went was said never for said non poked said non said non said non wasn't said non would wasn't poked never said would was something Murphy's said something for poked would found for found went non said would for said would for said poked never.
  Ma wasn't went was said never for said non poked said non said non said non wasn't said non would wasn't poked never said would was something Murphy's said something for poked would find for found went non said would for said poked never.
  BLEU 0.88, ROUGE 0.95
- themselves found, for apart about as were still in use on it get by up are to if can in should better as, were up and are to it get during in able Communists obliged, from, another if with solidly insufficient, still.
  They found, for apart about as were still in use on it get by up are to if can in should better as, were up and are to it get during in able Communists obliged, from another if solidly insufficient.
  BLEU 0.87, ROUGE 0.96

Phrases using Brown 3-gram model:

- Virginia and Rachel, conceded to be refined and dyed by the administration on the Corbin affair, as a board.
  Virginia and Rachel, conceded to be refined and dyed by the administration on the Corbin affair, as a board.
  BLEU 1.0, ROUGE 1.0
- fighters association here offered a $5,000 reward for information as to the wintry homeland of his office.
  The fighters association offered a $5,000 reward for information as to the winter homeland of his office.
  BLEU 0.71, ROUGE 0.89
- enemy in Korea and to organize reformatory occupants, defendants out on a green, the direction from which she poured turpentine.
  The enemy in Korea and to organize reformatory occupants, defendants out on a green, the direction from which she poured turpentine.
  BLEU 0.96, ROUGE 0.98
- floor whenever there was the land in the creche.
  When there was land in the creche, there was a floor.
  BLEU 0.25, ROUGE 0.6
- such as the believer lives under the will that is how to advance by increased allowances or new fund-raising efforts.
  Such as the believer lives under the will that is how to advance by increasing allowances or new fund-raising efforts.
  BLEU 0.81, ROUGE 0.95

Phrases using Brown 4-gram model:

- match, and at the same time upgrading quality.
  Match, and at the same time upgrading quality.
  BLEU 0.88, ROUGE 1.0

- and taste of any white wine will die a lingering death when it is directed by a leading merchant in Strasbourg whom she had already revealed a trill almost unprecedented in years of performances of
  And taste of any white wine will die a lingering death when it is directed by a leading merchant in Strasbourg, whom she had already revealed a trill almost unprecedented in years of performances of white wines.
  BLEU 0.82, ROUGE 0.97

- giving them an answer, I'm confident, because he always felt that he merely belonged among the myriad citizens of our community who are mentally unhinged
  I'm confident in giving them an answer because he always felt that he merely belongs among the myriad citizens of our community who are mentally unhinged.
  BLEU 0.70, ROUGE 0.83

- I find the performance less exciting than New York Democrats may wish, it nevertheless must be made.
  I find the performance less exciting than New York Democrats may wish, but it nevertheless must be made.
  BLEU 0.86, ROUGE 0.97

- were found in three regions of the body was buried under the kitchen floor or as dots posted over period marks in used books.
  They were found in three regions of the body, it was buried under the kitchen floor or as dots posted over period marks in used books.
  BLEU 0.96, ROUGE 0.98