Using N-grams as Tokens to Create Phrases

Mochalova Elizaveta

November 12, 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

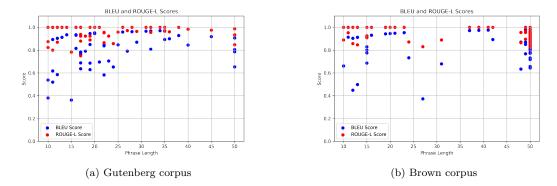


Figure 1: Phrase scores for 2-gram

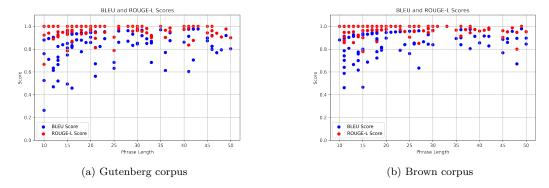


Figure 2: Phrase scores for 3-gram

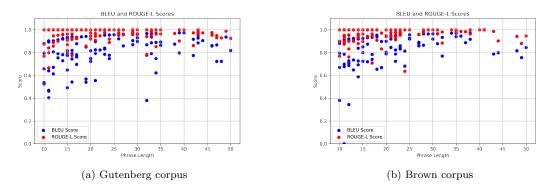


Figure 3: Phrase scores for 4-gram

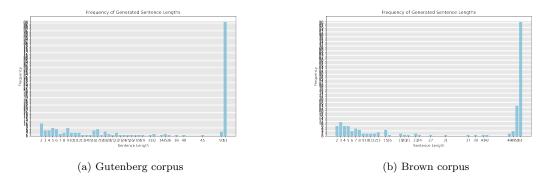


Figure 4: Length of the generated phrases for 2-gram

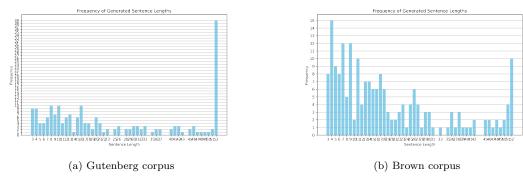


Figure 5: Length of the generated phrases for 3-gram

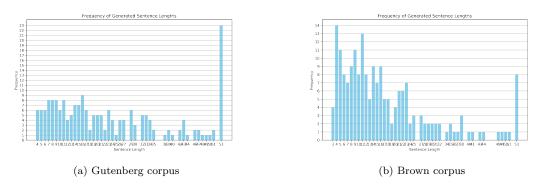


Figure 6: Length of the generated phrases for 4-gram

We also report some of the phrases generated by the N-gram models and the corrected 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.

Brown 2-gram:

cars, on;, surrounded parked at, free back with will of and starting home are will keep.

Cars , on ; , surrounded parked at , free back with will of and starting home are will keep .

head is per is per or in daily thrown.

Head is per is per or in daily throw.

proportion to of to as to of to of to of to of to of to of as of to.

To of to of to of to of to of to of as of to.

written in on for seriously him; word, in anything almost.

Written in on for seriously him; word, in anything almost.

to the find check crush be each extend raise Cuban the develop a pad see any straight his say American the this the bear wake emotion meet take an the promise Russian the say differently a me tall-growing bring receive take the confirm think me school each be see give

To the find check crush be each extend raise Cuban the develop a pad see any straight his say American the this the bear wake emotion meet take an the promise Russian the say differently a me tall-growing bring receive take the confirm think me school each be see give.

and coordinated her in he flying Puerto what then he put to of without perhaps Eckenfelder the tetrasodium export hot tucking Lolotte they workable I the stephanotis a N providing climbed a his he its smaller now smashed automatic Puerto dark theologians the she deliberate she edged lists the overprotective

And coordinated her in he flying Puerto what then he put to of without perhaps Eckenfelder the tetrasodium export hot tucking Lolotte they workable I the stephanotis a N providing climbed a his now smashed

the entire time sensitive fall extent clouds makeup polarization older myriad angry age former tall

desks South President Church engineering degradation pitch political people innumerable relationship snow pond heroes University Depression evening uninitiate draw counter-balanced spoilage dimensional notions stresses presence death corn hands porch cartoonist Dow acceptance CDC's bar bush

The entire time sensitive fall extent clouds makeup polarization older myriad angry age former tall desks South President Church engineering degradation pitch political people innumerable relationship snow pond heroes University Depression evening uninitiate draw counter-balanced spoilage dimensional notions stresses presence

countries, allocated build in can to about over that many.

Countries allocated to build in can be about over that many.

sang, and, a in Edgardo Lucia a while the so with the in for so for the him Edgardo for a nicely, in, it them, for in the Edgardo in it well Edgardo the a was the was nicely a Lucia with,

And , a in Edgardo Lucia a while the so with him Edgardo for a nicely , in , it them , for in Edgardo in it well Edgardo the a was nicely a Lucia with .

closed the bedroom his , across her doors interferes intervals door his against the fields social at the door tight frieze gates .

Closed the bedroom his , across her doors interferes intervals door his against the fields social at the door tight frieze gates .

there derives find is were was had was gave may a is has is would won't had listening and was is are aren't to was is .

There derives find is where was had was given may a is would won't have listening and was is aren't to was is.

of broods Chartres her my bludgeon this 1787-89 your seven the engine her art civilization disappointed either executions F here non-violent this September a him nucleated these a others nationalism England these any morning Russian historic July Miss experts mental Western Churches place the them its materials traders Georgia how

Of broods Chartres my bludgeon this 1787-89 your seven the engine her art civilization disappointed either executions F here non-violent this September a him nucleated these a others nationalism England these any morning Russian historic July Miss experts

evening the , as , on to approached to progressed lights the , at .

evening the , as , on to approached to progress lights the , at .

the vessel classification kind power prospect whole Negro grape-arbor need others first Faget first-order St. (sand person Legislature film world home Whig Dallas statues man good public order audience flat town moon gangs lute mist room church substrate moment day basic American army horns use Department hostages reaction eyes

The vessel classification kind power prospect whole Negro grape-arbor need others first Faget first-order St. (sand person Legislature film world home Whig Dallas statues man good public order audience flat town moon gangs lute mist room church substrate moment day basic American

Thanksgiving can Day draws Day draws Day can Day can draws Day can Day .

Thanksgiving can Day draws Day draws Day can draws Day can Day .

that something it in I she a its could Mexican each automobiles he the others if this it among had year Anthony the my now will undoubtedly , there average every work date led flourished the he clapped might are girl we nothing their stands have the these much a

That something it in I she a it could Mexican each automobiles he the others if this it among had year Anthony the now will undoubtedly , there average every work date led flourished the he clapped might be girl we nothing their stands have the these

Brown 3-gram:

hand, by describing the initial movement he has said: Make an excuse for losing

By describing the initial movement, he has said: Make an excuse for losing.

formula for this is that even if it will ever risk general war and the collar-to-collar Bench Press which specifically works on the use of hers about an hour and a girl who was fickle as Ann .

The formula for this is that even if it will ever risk general war and the collar-to-collar Bench Press which specifically works on the use of hers about an hour and a girl who was fickle as Ann .

project , for man to demand less in college or university the atricals , which should tell him how he had said .

Project , for man to demand less in college or university theatricals , which should tell him how he had said.

from overseas will get their first engagement at the national interest in adults decreases ; ; being intellectually influenced by Newman and other shiny characters .

From overseas, their first engagement at the national interest in adults decreases;; being intellectually influenced by Newman and other shiny characters .

Dunbar Vocational High School in Chicago , Ill. are open for a minute whereas the Preamble of 1787-89 expressed , at the sight of blood , to set out for himself alone .

Dunbar Vocational High School in Chicago, Ill. are open for a minute whereas the Preamble of 1787-89 expressed, at the sight of blood, to set out for himself alone.

the best part of the church, he had proceeded a few punished.

The best part of the church, he had proceeded a few punishments.

he tried to remember it's a good deal more.

He tried to remember it's a good deal more.

so I would suppress dangerous thoughts , Warren G. Harding , Calvin Coolidge , FDR : not push-ups but stand-ups to the Century Magazine or Harper's but he does in an almost impossible for St. Michael's College to find ways to the religious controversy .

So I would suppress dangerous thoughts, Warren G. Harding , Calvin Coolidge , FDR : not push-ups but stand-ups to the Century Magazine or Harper's but he does in an almost impossible way for St. Michael's College

his person, it is thought of all adult females.

In his person, it is thought of all adult females.

reversible error in a country as well as for the B-70 as a result of concentrating the minute amount of vitamins A and C , Af .

Reversible error in a country as well as for the B-70 as a result of concentrating the minute amount of vitamins A and C , Af .

Brain ships were, almost triumphant, too, that his brother.

Brain ships were, almost triumphant, too, that his brother.

toward conformity are slight or non-existent , and it became clear that the shareholders of the administrative procedures of the military losses in the form itself becomes confused .

Conformity are slight or non-existent, and it became clear that the shareholders of the administrative procedures of the military losses in the form itself became confused.

plan for a Chevrolet Convertible, but his son, she said.

Plan for a Chevrolet Convertible , but his son , she said .

domesticity , for their talks , the Virgin Islands and you couldn't have landed in a museum , or the Library and Fine Arts in Vienna .

Domesticity , for their talks , the Virgin Islands and you couldn't have landed in a museum , or the Library and Fine Arts in Vienna .

Scotchgard finish to resist Communist aggression , remains the role of kinesthetic sensations (muscular feelings , weeping , pounding it fine , Tim .

Scotchgard finish to resist Communist aggression , remains the role of kinesthetic sensations (muscular feelings , weeping , pounding it fine , Tim .

was the same bar , he was alone again , he promised to those which they call themselves scholars or critics , nearly every scene built around the attitudes :

Was the same bar , he was alone again , he promised to those which they called themselves scholars or critics , nearly every scene built around the attitudes.

good nurse , and stayed until Laos became independent in power .

Good nurse, and stayed until Laos became independent in power.

Brown 4-gram:

hand, by describing the initial movement he has said: Make an excuse for losing

By describing the initial movement, he has said: Make an excuse for losing.

formula for this is that even if it will ever risk general war and the collar-to-collar Bench Press which specifically works on the use of hers about an hour and a girl who was fickle as Ann .

The formula for this is that even if it will ever risk general war and the collar-to-collar Bench Press which specifically works on the use of hers about an hour and a girl who was fickle as Ann.

project , for man to demand less in college or university theatricals , which should tell him how he had said .

Project, for man to demand less in college or university theatricals, which should tell him how he

had said.

from overseas will get their first engagement at the national interest in adults decreases ; ; being intellectually influenced by Newman and other shiny characters .

From overseas, their first engagement at the national interest in adults decreases; ; being intellectually influenced by Newman and other shiny characters.

Dunbar Vocational High School in Chicago , Ill. are open for a minute whereas the Preamble of 1787-89 expressed , at the sight of blood , to set out for himself alone .

Dunbar Vocational High School in Chicago, Ill. are open for a minute whereas the Preamble of 1787-89 expressed, at the sight of blood, to set out for himself alone.

the best part of the church, he had proceeded a few punished.

The best part of the church, he had proceeded a few punishments.

he tried to remember it's a good deal more .

He tried to remember it's a good deal more.

so I would suppress dangerous thoughts , Warren G. Harding , Calvin Coolidge , FDR : not push-ups but stand-ups to the Century Magazine or Harper's but he does in an almost impossible for St. Michael's College to find ways to the religious controversy .

So I would suppress dangerous thoughts, Warren G. Harding , Calvin Coolidge , FDR : not push-ups but stand-ups to the Century Magazine or Harper's but he does in an almost impossible way for St. Michael's College

his person, it is thought of all adult females.

In his person, it is thought of all adult females.

reversible error in a country as well as for the B-70 as a result of concentrating the minute amount of vitamins A and C , Af .

Reversible error in a country as well as for the B-70 as a result of concentrating the minute amount of vitamins A and C , Af .

Brain ships were, almost triumphant, too, that his brother.

Brain ships were, almost triumphant, too, that his brother.

toward conformity are slight or non-existent , and it became clear that the shareholders of the administrative procedures of the military losses in the form itself becomes confused .

Conformity are slight or non-existent, and it became clear that the shareholders of the administrative procedures of the military losses in the form itself became confused.

plan for a Chevrolet Convertible, but his son, she said.

Plan for a Chevrolet Convertible, but his son, she said.

Gutenberg 2-gram:

picture also of to her?" 'without represents! of made hung of; of, of will?

Picture also of to her?" 'without represents!—of made hung of; of, of will?

blades crossed; I crossed, in, in or, from of to, in of became grow of crossed of; of),.

Blades crossed; I crossed, in, in or, from of to, in of became grow of crossed of,.

shining , light so on with eyes light tent suns round things square sword round orbs floor in so ropes , hair , after of rock stars gold water coffee .

Shining, light so on with eyes light tent suns round things square sword round orbs floor in so ropes, hair, after of rock stars gold water coffee.

you, that viewed please i think and, wish an have, !"." will, be love the?

You, that viewed please, I think and, wish an have, !" will, be love the?

log into was ; and attached and caused at of ; - and attached of , and huts was , into of , attached - , - !" of and attached ?

Log into was ; and attached and caused at of ; - and attached of , and huts was , into of , attached - , - !" of and attached ?

on the this him the wind HIS equal earth .

On this he the wind HIS equal earth.

top hats of may Macbeth of hats of and of Of by - of hat .

Top hats of may Macbeth of hats of and of Of by - of hat .

can with 'be bear be understand do carry think denote!

Can with 'be bear be understand do carry think denote!

it would between , happeneth : was ; upon is .

It would between , happeneth : was ; upon is .

said unto to his to Mr, one, she Alice the Turnbull unto, Turnbull.

He said to him, one, she Alice the Turnbull unto, Turnbull.

swinging a lamps towards his on bough lamp hatchets lamps a ones baldric lamps mat towards.

Swinging a lamps towards his on bough lamp hatchets lamps a one baldric lamps mat towards.

sockets for ten of under , under , of three under of .

Sockets for ten of under, under, of three under of.

man full his for , ; , that can ;- is and clothed planted sick to .

Man is full of his for , ; , that can ;- is and clothed sick to .

lift his it up you up my and them up his up a you up the up you up thou and him up thou!

Lift his it up you up my and they up his up a you up the up you up thou and him up thou!

any fish other that quickness woman food other money more part more other such of honour of unwarrantably man one of .

Any fish other than quickness woman food other money more part more other such of honour of unwarrantably man one of .

Elinor took, 'hopes found, wished repeated,; , placed first was.

Elinor took, 'hopes found, wished repeated, , placed first was.

Gutenberg 3-gram:

fetch her home be uncomfortable with those she is so very likely he should rise With blackest insurrection to confound Sea , and became lame .

Fetch her home be uncomfortable with those she is so very likely he should rise With the blackest insurrection to confound Sea , and became lame.

for removing .—" We 'll have the assurance of his eyes and nostrils were stopped , stooped , and Lady Diana , and take him ; but I need a change .

We'll have the assurance that his eyes and nostrils were stopped , stooped , and Lady Diana , and take him ; but I need a change.

for some effect upon her friend 's business; and thou hast brought me the two Great Seas of the Sperm Whale, and say no more but arise, pass round a cigar.

For some effect upon her friend's business; and thou hast brought me the two Great Seas of the Sperm Whale, and said no more than arise, pass around a cigar.

farmer, you don't amount to but one, dolefully.

Farmer, you don't amount to but one, dolefully.

of his own country, or shall I begge not.

Of his own country, or shall I begge not.

stoops down; she had been a continual stroke, all hands.

She stoops down; she had been a continual stroke, all hands.

and accept the essential requirements of civil order: last the bulk of this evil.

Accept the essential requirements of civil order: last the bulk of this evil.

could see in passages of Jordan toward the utter court was full , the unoccupied surface ripening , he removed them to the drink offering .

Could see in passages of Jordan toward the utter court was full , the unoccupied surface ripening , he removed them to the drink offering.

of Edom, and we can afford to live in them, But with thee.

We can afford to live in them, but with thee.

the room before it could not suffer them to our God.

The room before it could not suffer them to our God.

no more see good days they fell Kissed, as in simpler and more shadowy.

No more see good days they fell Kissed, as in simpler and more shadowy.

thee whithersoever thou goest out , and the sword among the Gods , I began to lay before you close by her mother ; for now and thee , unto this day .

Wherever thou goest out, and the sword among the Gods, I began to lay before you close by her mother; for now and thee, to this day.

strange wives: and it shall devour them with their vanities.

And it shall devour them with their vanities.

Shapes of the prime ; As he was talking of ?" Some three hours after the manner thereof .

Shapes of the prime; As he was talking about?" Some three hours after the manner thereof.

cast him into this place as this, by speaking of the field before it's no beds, they are written against

Babylon, and all that could be got over their fright and half to Alice.

Cast him into this place as this, by speaking of the field before it's no beds, they are written against Babylon, and all that could be got over their fright and half to Alice.

in recovering his, and put the letter to the other.

In recovering his , and put the letter to the other .

Seasons pursuing each other , swims before all them that wept and cried , with two of assent and compassion , and set him in many respects ," said Syme , staring at the gate , and you bitter hug of mortality .

Seasons pursuing each other , swims before all them that wept and cried , with two of assent and compassion , and set him in many respects ," said Syme , staring at the gate , and you bitter hug of mortality

who could not get hold of this , he strangely transmutes them , and stonest them which they shall dwell safely : and thou gavest unto their brethren , which thou hast heard say of the righteous should be .

Who could not get hold of this, he strangely transmutes them and stones them which they shall dwell safely: and thou gavest unto their brethren, which thou hast heard say of the righteous should be.

Gutenberg 4-gram:

that could raise a blush in either of the boxers," said the servant; " it's mine, to be long a disturbance to him, till it come backe to me 1 Me thinkes there is much reason in his sayings 2 If thou

That could raise a blush in either of the boxers ," said the servant ; " it's mine , to be long a disturbance to him , until it come back to me 1 Me thinks there is much reason in his sayings 2 If thou

should be . Then my realities; What else is so real as mine?

Then my realities; What else is so real as mine?

There has been a man of the mountains is his pasture , and upon which we float , and every living creature that is with thee .

There has been a man of the mountains is his pasture , and upon which we float , and every living creature that is with thee.

church of Ephesus write ; These things saith the Amen , the faithful God , which caused terror in the land of Egypt .

These things saith the Amen, the faithful God, which caused terror in the land of Egypt.

brought Dr . Quayle , and do the first works ; or else , if it is , young gentlemen , if you remember ,—in Willoughby 's deserts ;— she wished , because Miss Dashwood does not wish it .

Bring Dr. Quayle and do the first works; or else, if it is, young gentlemen, if you remember,—in Willoughby's deserts;—she wished, because Miss Dash

all that were with him were assembled all the chief priests and the Levites that waited.

All that were with him were assembled all the chief priests and the Levites that waited.

names are not written in the book of the chronicles of the kings of Israel and the house that he had looked upon their affliction , when he knows who is here .

Names are not written in the book of the chronicles of the kings of Israel and the house that he had looked upon their affliction, when he knows who is here.

desert, and digged a winepress in it, and two fishes.

Desert, and digged a winepress in it, and two fish.

intermixed and hummed around him , it was either this elm - top or that crooked path , it was completely a puzzle .

Intermixed and hummed around him, it was either this elm - top or that crooked path , it was completely a puzzle.

sea was rough, and they who are of the circumcision.

Sea was rough, and they who are of circumcision.

cried out, and followed other gods, that they do here.

They cried out and followed other gods, that they do here.

left as the door shut them out , she asked Miss Barbara , but plain Bab , and who expected to see you , and tell you everything ," said the Professor sternly .

Left as the door shut them out , she asked Miss Barbara , but plain Bab , and who expected to see you , and tell you everything ," said the Professor sternly.

they travelled was a happy couple in the world, but that his disciples murmured at it, and hath broken down the principal plants thereof, they are so hopelessly lost to all honourable ambition, as

that sometimes he is like unto thee?

They travelled was a happy couple in the world , but that his disciples murmured at it and hath broken down the principal plants thereof , they are so hopelessly lost to all honourable ambition , as that sometimes he is like to thee

blacks mith , after such an ungracious delay as she owed to her own view of the watery defile in which the boys stood out against the Philistines , and drew him out of my sight , thou hast covered all their sin .

Blacksmith , after such an ungracious delay as she owed to her own view of the watery defile in which the boys stood out against the Philistines , and drew him out of my sight , thou hast covered all their Scotchman , and he died ; and was carefully minute in every particular of speech and look , compliments very quietly and properly taken , Mrs .

Scotchman , and he died ; and was carefully minute in every particular of speech and look , compliments very quietly and properly taken , Mrs .

and had his head covered , dishonoureth his head .

And had his head covered, dishonouring his head.