Neural Unsupervised Paraphrasing

Team Name : **Confidant-tri-gram**

Zeeshan Khan (2021701029)
K. N. Amruth Sagar (2021701013)
Seshadri Mazumder (2021801002)

Course Name : Introduction to NLP
Course Code : CS7.401
Semester : Spring'22
Instructor Name : Prof. Manish Srivastava
Assigned TA : Ananya Mukherjee

International Institute of Information Technology
Hyderabad, India

May 1, 2022

Project Link

Contents

1 Abstract

Neural unsupervised Paraphrasing is one of the interest of the recent NLP researchers. This research helps to use datasets which doesn't have human annotations and thus opens a new direction in the field of paraphrasing. We have use pretrained BART to form pseudo Ground truth and then perform an end to end supervised training of transformers from scratch. We used four different corruption techniques and that makes our model more robust.

2 Introduction

Paraphrasing is rewriting a given sentence in other words or forms without losing the meaning of the original sentence. In NLP, there are two types of problems that involve paraphrasing, paraphrase generation, and paraphrase detection. Paraphrase detection determines whether a given pair of sentences have the same meaning or not. In contrast, paraphrase generation will generate a sentence with different words but the same meaning as the given input sentence.

Paraphrase generation can be done through various approaches. Few approaches use machine translation as an intermediate step to produce the paraphrases from the translated sentence. Some use a supervised system by training the models with a large parallel corpus. Some make paraphrases using optimization techniques like simulated annealing. Others follow unsupervised approaches to train models.

This project is about unsupervised paraphrasing.[1] Supervised methods require extensive labeled data, which is hard to collect, making the unsupervised approach a better alternative. In this project, we take a pre-trained language model and use masking strategies to corrupt the sentence and generate sentences, which form pseudo paraphrase pairs, which form the self-supervision data, and then train the language using this data.

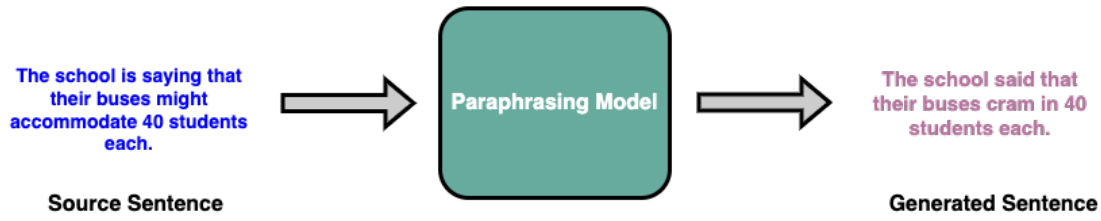

Figure 1: Problem Statement

Given a sentence $S = \{s_1, s_2, ..., s_n\}$ from a vocab domain V, we will learn a transformation function P such that, $P(S) = T$, where $T = t_1, t_2, ..., t_n$, such that $\{t_1, t_2, t_n\} \in V$ and sentence T is a paraphrased sentence of S having consistency in the semantic domain.

[1]Project link

3 Related Works

Paraphrasing is a well-known problem in natural language processing (NLP), with numerous applications such as data augmentation, data curation, intent mapping, and semantic comprehension. Paraphrasing can be used to create synthetic data to supplement sparse datasets for training, as well as to improve the performance of downstream tasks like classification. Because of the linguistic heterogeneity added by paraphrasing, it can also be utilised to extend current datasets to more variations, making them more generalizable.

The task of textual paraphrasing has been tackled using variety of approaches. There has been a lot of work on framing generation of paraphrases as a Seq2seq task. Previous explorations have largely focused on supervised methods, which require a large amount of labeled data that is costly to collect. So unsupervised approaches are suitable in this case as it requires huge effort to prepare a paraphrase dataset.

Some unsupervised approaches employ Deep reinforcement learning, in which a variational autoencoder trained on a non-parallel corpus is used to create a seed para that warms up the DRL model, and then the seed para is fine-tuned over time. [1]

Other methods recast the problem as an optimization problem with a complex objective function. Then, using techniques like simulated annealing, a solution is found in the sentence space based on the locally conducted operations on the sentences. [2]

Few methods do paraphrasing without the need for machine translation as an intermediary step in the paraphrase creation process. [3]

The approach we are interested is where we leverage pretrained language models, because they are trained on huge corpus, and are very good at predicting the next word in the sequence. They can be fine tuned for the paraphrase generation task, where the generated sentences can be used to create a parallel dataset, which can then be used to train a model using it. [4]

4 Methodology

4.1 Dataset

4.1.1 QQP : Quora Question Pair dataset

Quora Question Pairs (QQP) dataset consists of over 400,000 question pairs, and each question pair is annotated with a binary value indicating whether the two questions are paraphrase of each other.

- Total no of rows in data = **404290**
- There exists no Test or train split
- No of words in vocab
  Total = 73948

- QQP

4.1.2 PAWS : Paraphrase Adversaries from Word Scrambling dataset

The PAWS dataset contains 108,463 human-labeled and 656k noisy labelled pairs that demonstrate the relevance of modelling structure, context, and word order information for the challenge of paraphrase detection. The dataset is divided into two parts, one based on Wikipedia and the other on the Quora Question Pairs (QQP) dataset.

We considered the Wikipedia part of the dataset and in that, we choose PAWS-wiki-final, which contains pairs created using both word swapping and back translation techniques, along with human judgements on both paraphrasing and fluency.

- Total no of rows in data (train = 49401 + test = 8000 + dev = 8000) = **65401**
- The data is divided into test, train and dev
- No of words in vocab
  Train = 29827, Test = 8704, Dev = 8818, All = 35583
- PAWS

4.1.3 Microsoft Research Paraphrase Corpus

Microsoft Research Paraphrase Corpus (MRPC) is a corpus consists of sentence pairs collected from newswire articles. Each pair is labelled if it is a paraphrase or not by human annotators.

- Total no of rows in data (train + test) = **5803**
- There exists a Test and a Train split
- No of words in vocab
  Train = 21608, Test = 12237, Train+Test = 21608
- MRPC

4.2 Pipeline

In the first stage, every sentence of the source sentences is corrupted in 4 ways, and we get four different corrupted sentences for each sentence. And then, these sentences are passed through pretrained BART, and the model will try to fill in the masked positions present in the input and generate sentences.
These generated sentences are treated as the target/paraphrase part of a sentence-paraphrase pair, and a parallel dataset is prepared by concatenating the source sentences and the generated sentences. Then this dataset is used to train a Seq2Seq Transformer model in a supervised fashion.

4.3 Data Corruption

- **Method 1 : Token masking**

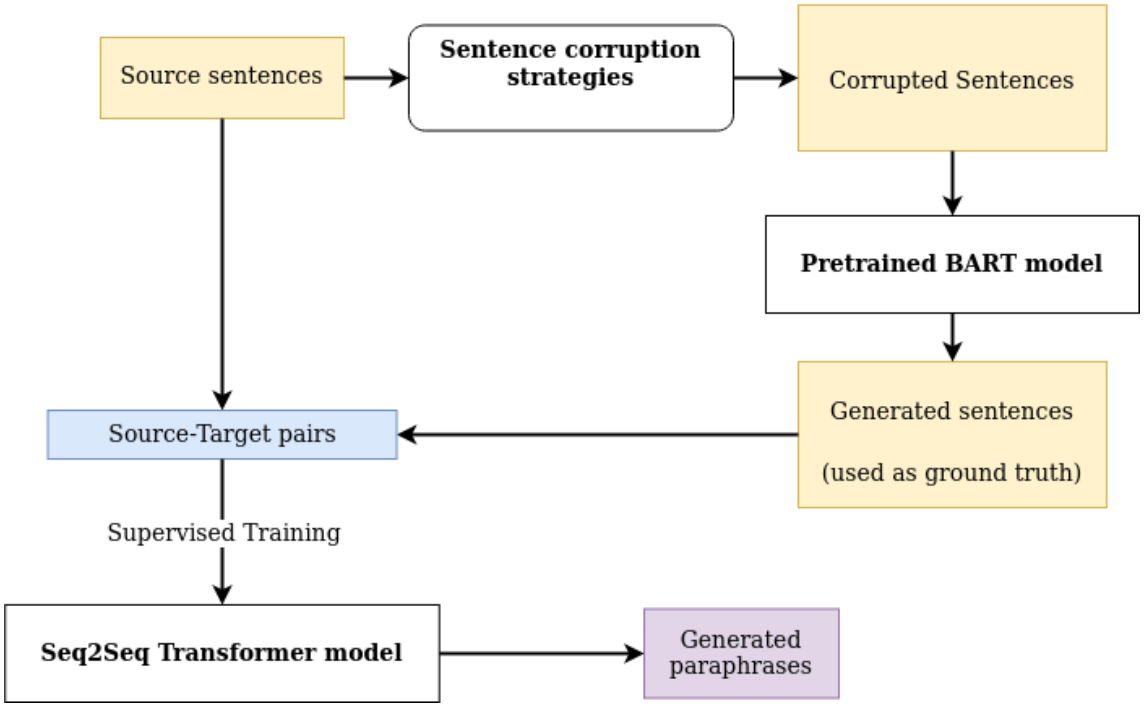

Figure 2: Pipeline

In token masking, a few tokens are randomly selected and a are replaced with the <mask> tag.

Examples : The first line represents source and the following line represents the predictions/outputs.

```
Amrozi accused his brother, whom he called "the witness", of deliberately
distorting his evidence.
Amrozi accused his brother, <mask> he <mask> "the witness", of <mask>
distorting his evidence.

Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for
$2.5 billion.
Yucaipa owned Dominick's <mask> selling <mask> chain to Safeway <mask>
1998 for $2.5 billion.

They had published an advertisement on the Internet on June 10, offering
the cargo for sale, he added.
They <mask> published an advertisement <mask> the Internet on June <mask>
offering the cargo for sale, <mask> added.

Around 0335 GMT, Tab shares were up 19 cents, or 4.4%, at A$4.56, having
earlier set a record high of A$4.57.
Around <mask> GMT, Tab <mask> were up 19 cents, or 4.4%, at A$4.56, <mask>
earlier set a record <mask> of A$4.57.
```

- **Method 2 : Token deletion**

  In token deletion, a few tokens are randomly selected and then are removed from the sentence.

  Examples : The first line represents source and the following line represents the predictions/outputs.

  ```
  Amrozi accused his brother, whom he called "the witness", of deliberately
  distorting his evidence.
  Amrozi accused his brother, whom he called "the witness", deliberately
  distorting

  Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for
  $2.5 billion.
  owned Dominick's before the chain to Safeway 1998 for $2.5 billion.

  They had published an advertisement on the Internet on June 10, offering
  the cargo for sale, he added.
  had an advertisement on the Internet on June 10, offering the sale,
  he added.

  Around 0335 GMT, Tab shares were up 19 cents, or 4.4%, at A$4.56, having
  earlier set a record high of A$4.57.
  Around 0335 GMT, shares up 19 cents, or 4.4%, at A$4.56, having earlier a
  of A$4.57.
  ```

- **Method 3 : Text infilling**

  Instead of masking just the token at a randomly selected position, we will replace multiple tokens at the given position based on the span lengths, with `<mask>`. If the span length is > 0, those many tokens from that position are replaced with `<mask>`. If the span length is 0, then `<mask>` is inserted between two tokens.
  The span lengths for text infilling is sampled from a gamma distribution, and the chosen gamma value was 2.

  Examples : The first line represents source and the following line represents the predictions/outputs.

  ```
  Amrozi accused his brother, whom he called "the witness", of deliberately
  distorting his evidence.
  Amrozi accused his brother, whom he <mask> "the witness", of deliberately
  <mask> distorting his evidence. <mask>
  ```

```
Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for
$2.5 billion.
Yucaipa owned Dominick's before selling the chain to Safeway in 1998
<mask> <mask> <mask>

They had published an advertisement on the Internet on June 10, offering
the cargo for sale, he added.
They had published an advertisement <mask> the <mask> on June <mask> 10,
offering the <mask> for sale, he added.

Around 0335 GMT, Tab shares were up 19 cents, or 4.4%, at A$4.56, having
earlier set a record high of A$4.57.
<mask> 0335 GMT, Tab shares were up <mask> cents, or <mask> <mask> A$4.56, having
```

- **Method 4 : Document rotation**

  In document rotation, a token is randomly selected and the whole document(sentence in this case) is rotated such that the selected token is the first word of the sentence.

  Examples : The first line represents source and the following line represents the predictions/outputs.

```
Amrozi accused his brother, whom he called "the witness", of deliberately
distorting his evidence.
of deliberately distorting his evidence. Amrozi accused his brother, whom
he called "the witness",

Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for
$2.5 billion.
to Safeway in 1998 for $2.5 billion. Yucaipa owned Dominick's before
selling the chain

They had published an advertisement on the Internet on June 10, offering
the cargo for sale, he added.
published an advertisement on the Internet on June 10, offering the cargo
for sale, he added. They had

Around 0335 GMT, Tab shares were up 19 cents, or 4.4%, at A$4.56, having
earlier set a record high of A$4.57.
high of A$4.57. Around 0335 GMT, Tab shares were up 19 cents, or 4.4%, at
A$4.56, having earlier set a record
```

4.4 BART Inference

A pretrained BART model from hugging face is downloaded, and then the model is set to *eval()* mode. Then, the corrupted sentences are are fed to the model to fill in the mask with words, in other words, to reconstruct the original sequence before corruption.

DATASET	TIME
MSRP train.txt	$\approx$ 18 minutes
paws wiki train.txt	$\approx$ 3 hours
QQP dataset	$\approx$ 12 hours

Table 1: Table to test captions and labels.

4.5 Training Transformers from Scratch

4.5.1 Data Preparation

- **Preparing Source and Target Data**

  Write about the following things :
  - Cleaning the Sentence using Regular Expression : We have used Regular Expressions to clean the sentences and remove different stopwords.
  - Adding SOS and EOS and padding the sequence for a batch : After we clean the sentences we attach SOS & EOS tokens at the start and the end of the sentences respectively and then pad the sentences upto a same length.
  - Numericalizing the Sentence : Here we map the words of a sentence to vocab space where we have one to one mapping of word with an integer.

- **Preparing Source and Target Mask**

  Write about the following things :
  - Source Masking : Here we create an boolean array fopr TRUE where we have valid words ignoring the PAD tokens where we use boolean FALSE.
  - Target Masking : In target masking we use an nxn boolean matrix in which the lower diagonal of the matrix is filled with boolean TRUE and the rest are filled with boolean FALSE. This is to ensure that during testing the model, the model doesn't see the future words, so we just mask them.

4.5.2 Architecture Discussion

4.6 Design choices

- **Label Smoothing** : Label smoothing is used to eliminate overconfidence that usually arises during classifications tasks, especially in deep models. In label smoothing,

if we are using cross entropy as the loss function and if the model is applying softmax in the last layer to generate k classes, then instead of sending the one hot encoded vector $\hat{y}$, we will send a combination of one-hot and uniform distribution.

$$\hat{y}_{ls} = (1 - \alpha)\hat{y} + \alpha/k$$

Here, α is the super parameter. If $\alpha = 0$, then we get original one-hot, if $\alpha = 1$, then we get a uniform distribution.

- **Loss Function** : Kullback-Leibler divergence, also called as relative entropy, calculates a score that measures the divergence of one probability distribution from another.
- **Optimization** : *Noam optimizer* is used, which was also used in *Attention is all you need* paper. It is an Adam optimizer, with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-6}$. Initially, the learning rate is allowed to increase linearly in the *warmup* training stages, which is about 4000 steps, and then the learning rate decreases proportionally to the inverse square root of the step number.

4.7 Training Details

All the models were trained using 4 2080ti GPU's with `nn.DataParallel`

Dataset	Train epoch time	Valid epoch time	Batch size
MSR	59 seconds	21 seconds	64
PAWS	12.19 minutes	1.42 minutes	64
QQP	2.42 hours	-	64

Table 2: Training details

5 Results & Discussion

5.1 BART Inference Outputs

The first line represents source and the following line represents the predictions/outputs.

```
The NBA season of 1975 -- <mask> was the 30th <mask> of the <mask> Basketball
Association <mask>
The NBA season of 1975 -- 1976 was the 30th season of the National Basketball
Association

There are also <mask> discussions <mask> public profile <mask> and project
discussions .
```

There are also private discussions, public profile discussions, and project discussions.

When comparable rates of flow can be maintained , the results <mask> <mask> <mask>
When comparable rates of flow can be maintained, the results can be summarized as follows:

It is the seat <mask> <mask> District in <mask> Region .
It is the seat of the Mpumalanga District in Zambia's Eastern

5.2 Training Transformers on QQP dataset

QQP Dataset, we can only train for one epoch for due to time and resource constraint.

5.2.1 Quantitative Evaluation

In the first Epoch since we have aplied multiple corruption techniques and due to high amount of variances in data our model achieved 7.31 BLEU score.

5.3 Training Transformers on PAWS dataset

5.3.1 Qualitative Evaluation

Pair 1

Target : JOHN BARROW ISLAND IS A MEMBER OF THE QUEEN ELIZABETH ISLANDS IN THE CANADIAN ARCTIC ARCH
Predicted : JOHN A MEMBER OF PARLIAMENT OF THE CANADIAN PROVINCE OF CANADA IN THE CANADIAN PROVINCE

Pair 2

Target : BEST MOVIE AT THE 7TH YOKOHAMA FILM FESTIVAL IT WAS CHOSEN AS THE
Predicted : IT WAS THE FIRST TIME AS THE SECOND TIME AS THE 7TH BEST FILM FESTIVAL

Pair 3

Target : ALL FIVE EVENTS STARTED THE LAST DAY OF THE TOURNAMENT AND CONCLUDED WITH THE FINAL ON THE
Predicted : THE FIRST FIVE ON THE AREA WERE THE FIRST AND LAST WITH THE FINAL

5.3.2 Quantitative Evaluation

- **Test BLEU Score :** We achieved a BLEU score of 7.312 on test set of MSR Paraphrase dataset. This score is from checkpoint : 72.
- **Training & Validation Loss graph**

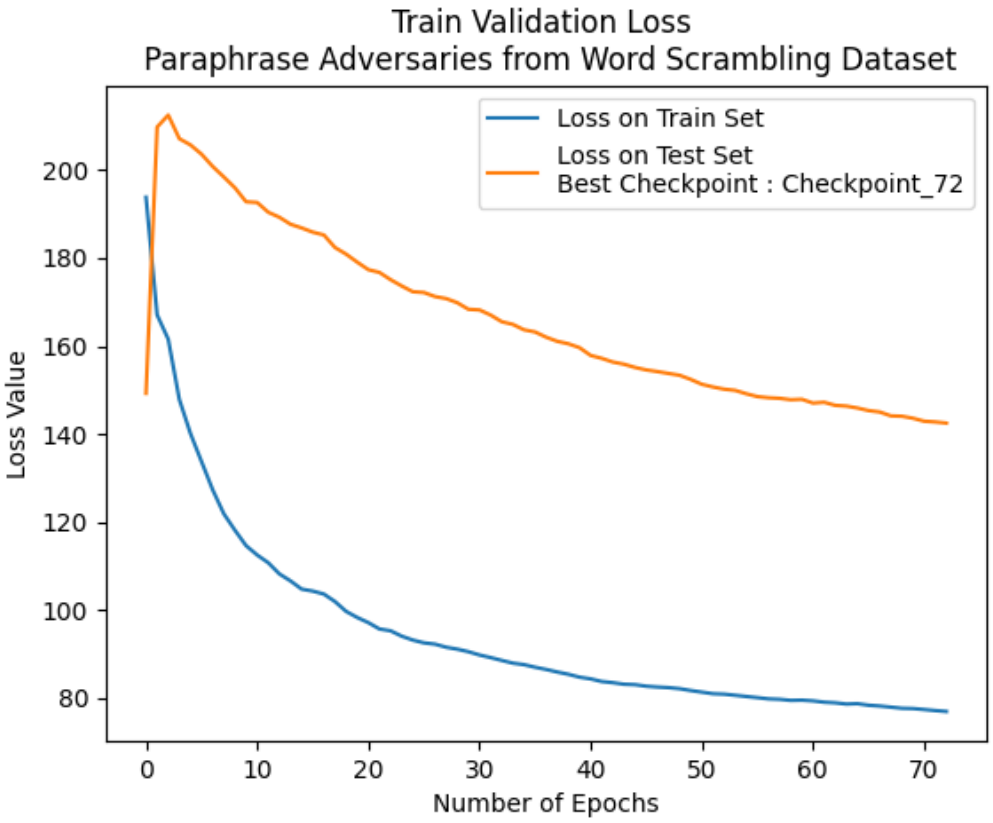

Figure 3: Training & Validation Loss Plot

5.4 Training Transformers on MSR paraphrase dataset

5.4.1 Qualitative Evaluation

```
    Pair 1 :

Target : THE WORLD S LARGEST AUTOMAKERS SAID THEIR U S SALES DECLINED MORE THAN
EXPECTED LAST
Predicted : THAN EXPECTED THE WORLD S TWO LARGEST AUTOMAKERS SAID THEIR U S SALES
DECLINED

Pair 2:

Target : LOSSES AND FAVORABLE INTEREST RATES THE INCREASE REFLECTS LOWER CREDIT
Predicted : THE INCREASE REFLECTS LOWER CREDIT LOSSES AND FAVORABLE MARKET
CONDITIONS

Pair 3 :

Target : FEDERAL AGENTS SAID YESTERDAY THEY ARE INVESTIGATING THEFT OF 1 200
OF AN EXPLOSIVE CHEMICAL FROM
Predicted : FEDERAL AGENTS SAID YESTERDAY THEY ARE INVESTIGATING THE THEFT OF
A TOTAL OF 1 200 POUNDS
```

5.4.2 Quantitative Evaluation

- **Test BLEU Score :** We achieved a BLEU score of 53.439 on test set of MSR Paraphrase dataset. This score is from checkpoint : 278.
- **Training & Validation Loss graph**

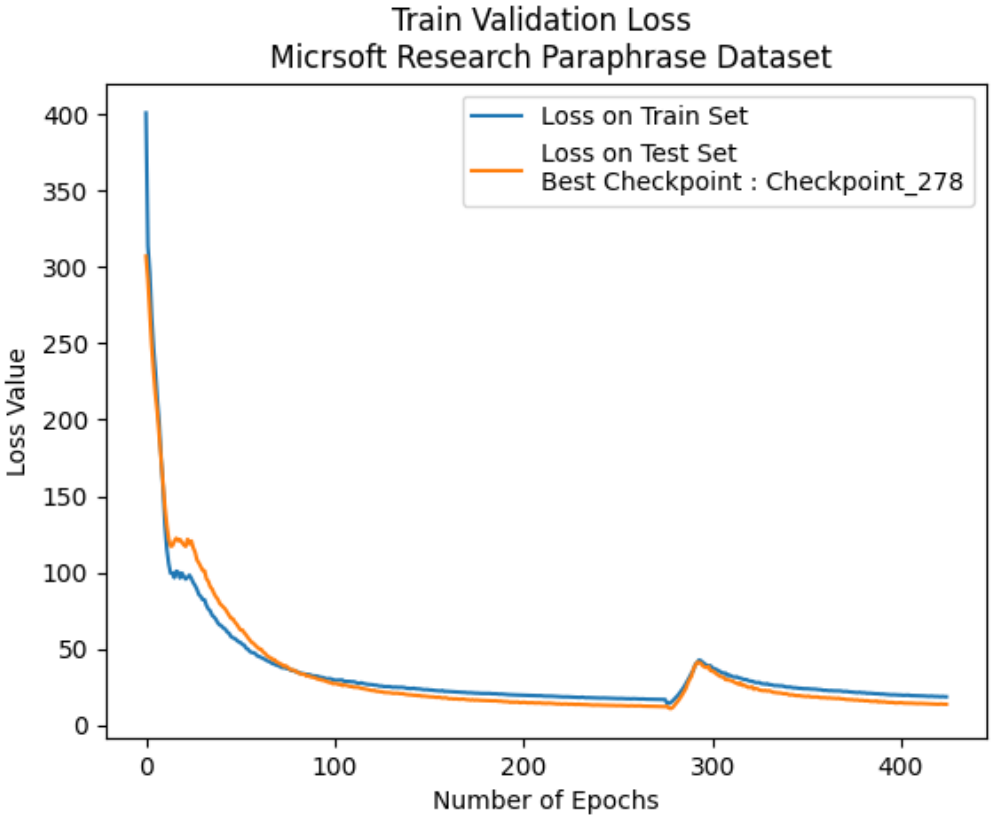

Figure 4: Training & Validation Loss Plot

6 Conclusion

Neural unsupervised is yet a challenging task and there's a lot of further research scope in this field. We first corrupt the sentences and then used BART's model to curate pseudo ground truth sentences. Then we trained an supervised end to end transformers from scratch and wqe achieve reasonable accuracies in the MSR dataset. We couldn't train our model on QQP dataset due to time & resource constraint. We also achieve some scores on the PAWS dataset, however the training loss graph shows, that further training can improve the results more, and there's a lot of scope for loss convergence. Experimenting with multiple datasets and jointly training them for various domains and using gradient blending between them still remains as a future scope to us.

References

[1] A. Siddique, S. Oymak, and V. Hristidis, "Unsupervised paraphrasing via deep reinforcement learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1800–1809, 2020.

[2] X. Liu, L. Mou, F. Meng, H. Zhou, J. Zhou, and S. Song, "Unsupervised paraphrasing by simulated annealing," *arXiv preprint arXiv:1909.03588*, 2019.

[3] A. Roy and D. Grangier, "Unsupervised paraphrasing without translation," *arXiv preprint arXiv:1905.12752*, 2019.

[4] C. Hegde and S. Patil, "Unsupervised paraphrase generation using pre-trained language models," *arXiv preprint arXiv:2006.05477*, 2020.