Neural Unsupervised Paraphrasing

Course Name : Intro to NLP
Course ID : CS7.401
Instructors : Prof. Manish Srivastava
Assigned TA : Ananya Mukherjee

Team: Confidant-tri-gram

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

Project Link : NLP_Project

Index

Introduction

- Paraphrasing is rewriting a given sentence in other words or forms without losing the meaning of the original sentence.
- Paraphrasing can be achieved via training a deep sequence to sequence models, like RNN, LSTMs, Transformers etc, in a supervised setting.
- Creating supervised datasets for paraphrasing is both time and cost expensive.
- Therefore, several approaches aims to achieve paraphrasing using unsupervised methods.
- In this project we aim at generating paraphrases using the idea of corrupting a sentence and reconstructing the correct sentence using deep auto-regressive denoising approaches.

Problem Statement

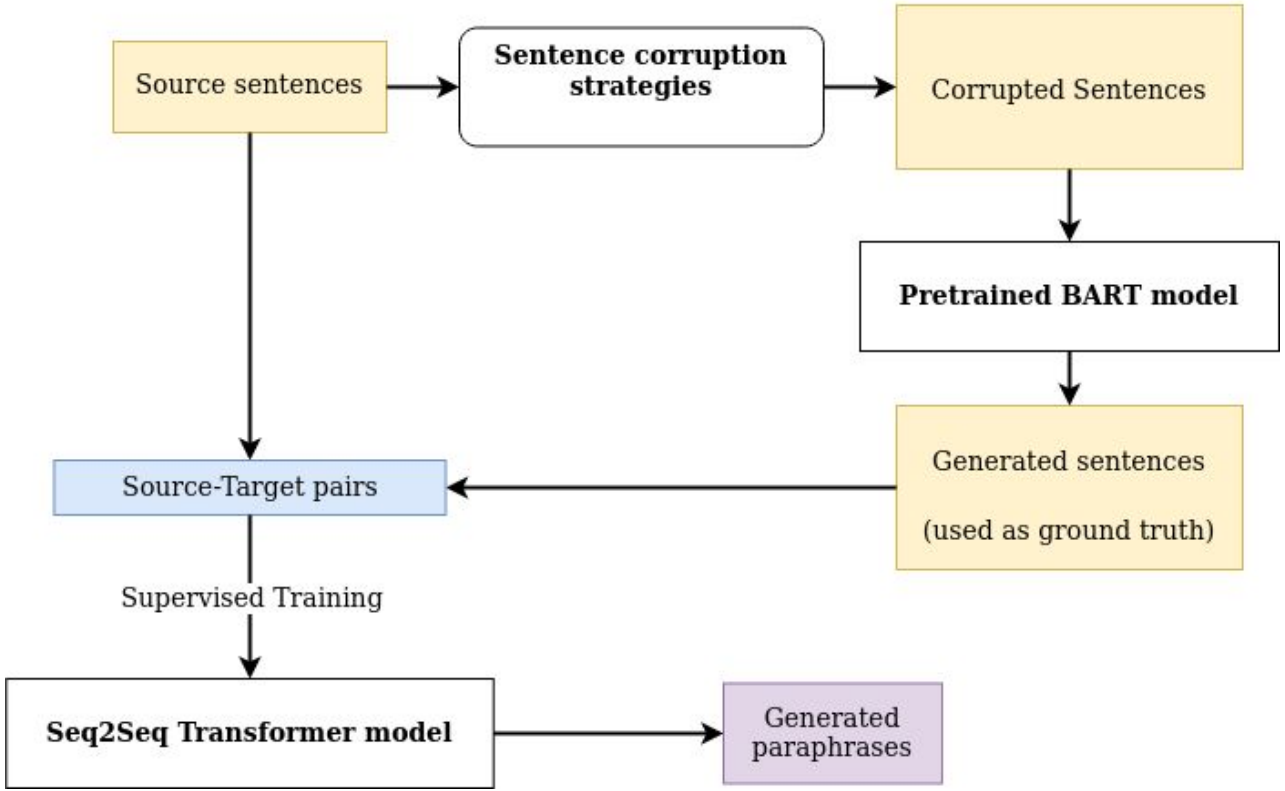
Given a sentence in natural language, we aim at generating a new sentence with the same meaning but different language structure in an unsupervised setting.

Datasets

We train and evaluate our model on 3 datasets:

Dataset	Train sentences	Valid sentences	Test Sentences	vocab
QQP : Quora Question Pair dataset	404290	-	-	73948
PAWS : Paraphrase Adversaries from Word Scrambling dataset	49401	8000	8000	35583
Microsoft Research Paraphrase Corpus	5803	-	-	21608

Pipeline


Source sentences
Sentence corruption strategies
Corrupted Sentences
Pretrained BART model
Generated sentences
(used as ground truth)
Source-Target pairs
Supervised Training
Seq2Seq Transformer model
Generated paraphrases

Data Corruption Techniques

- Token Masking:
  - Ex: **Source sentence**: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
  - **Corrupted sentence**: Amrozi accused his brother, <mask> he <mask> "the witness", of <mask> distorting his evidence
- Token Masking:
  - Ex: **Source sentence**: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
  - **Corrupted sentence**: Amrozi accused his brother, he called "the witness", of deliberately his evidence.

Data Corruption Techniques

- Text Infilling:
  - Ex: **Source sentence**: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
  - **Corrupted sentence**: Amrozi accused his brother, <mask> "the witness", of deliberately distorting his evidence
- Document rotation:
  - Ex: **Source sentence**: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
  - **Corrupted sentence**: Of deliberately his evidence. Amrozi accused his brother, he called "the witness".

Taking inference from BART Model

- Bart is an auto-regressive generative language model.
- It is pre-trained on the task of masking and autoregressive denoising.
- It has a transformer encoder which encodes a corrupted/masked sentence
- It has a transformer decoder which predicts the masked tokens in an autoregressive manner.
- Since we are working in an unsupervised setting, we create pseudo-ground truths for the task of paraphrase generation.
- We corrupt the sentences using the 4 strategies described previously.
- The noisy/masked sentence is fed to BART, it denoises the sentence and reconstructs it in an autoregressive manner.

What we have now ?

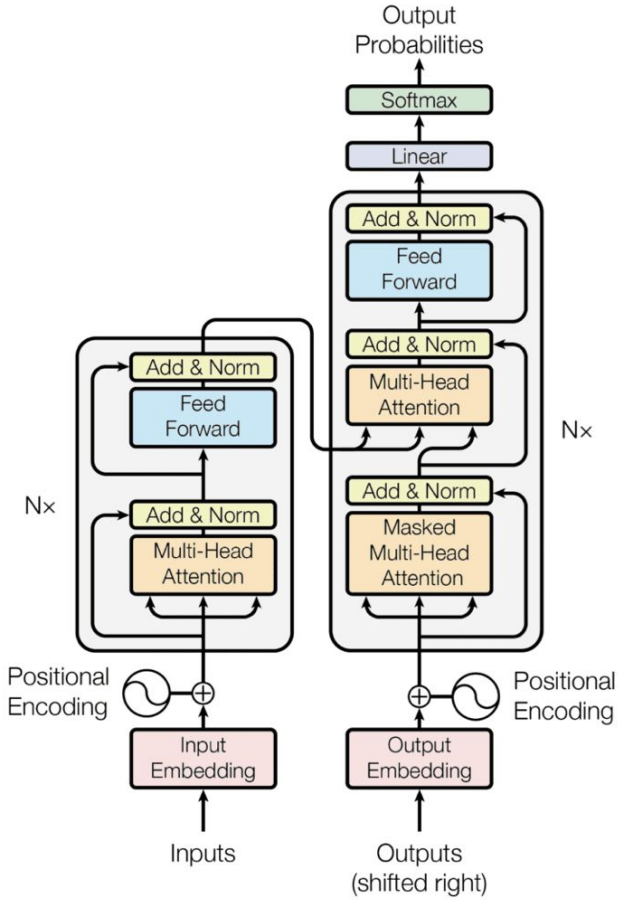
- Now we have a source sentence.
- 4 Corresponding masked/corrupted sentence.
- 4 Denoised sentences.
- We use the source sentence as the source.
- We use the 4 denoised sentences as pseudo-ground truths.

Transformers : Data Preparation

- We clean the sentences, numericalise the tokens and convert a sentence to a tensor.
- We Add <SOS> <EOS> tokens to represent start and end of statements.
- To take advantage of large scale training in batches which improves the transformers performance, we add the functionality of batching in our training.
- To enable batching we converted all the sentences to have a same length.
- For that we define a max length of tokens for any given batch of sentences, and add padding to fill in the max length. We use <PAD> token for padding.
- To handle the padding during training, we create transformer padding mask of size max length for the whole batch, with 1 where real token exists and 0 for <PAD> token.
- Since we use teacher forcing for autoregressive decoding, we create a transformer attention mask for the decoder, with upper triangular matrix filled with 1 and rest 0. This makes sure that the model does not look into the future tokens and cheat.

Standard Transformer Architecture

- Transformer is an encoder decoder style deep neural network model.
- It is able to parallely process the entire sequence while attending to the useful information using key, query and, value style attention and encode the whole sequence. It uses self attention in the encoder.
- In the decoder it uses cross attention to extract useful information from the encoded sequence to help with the decoding.



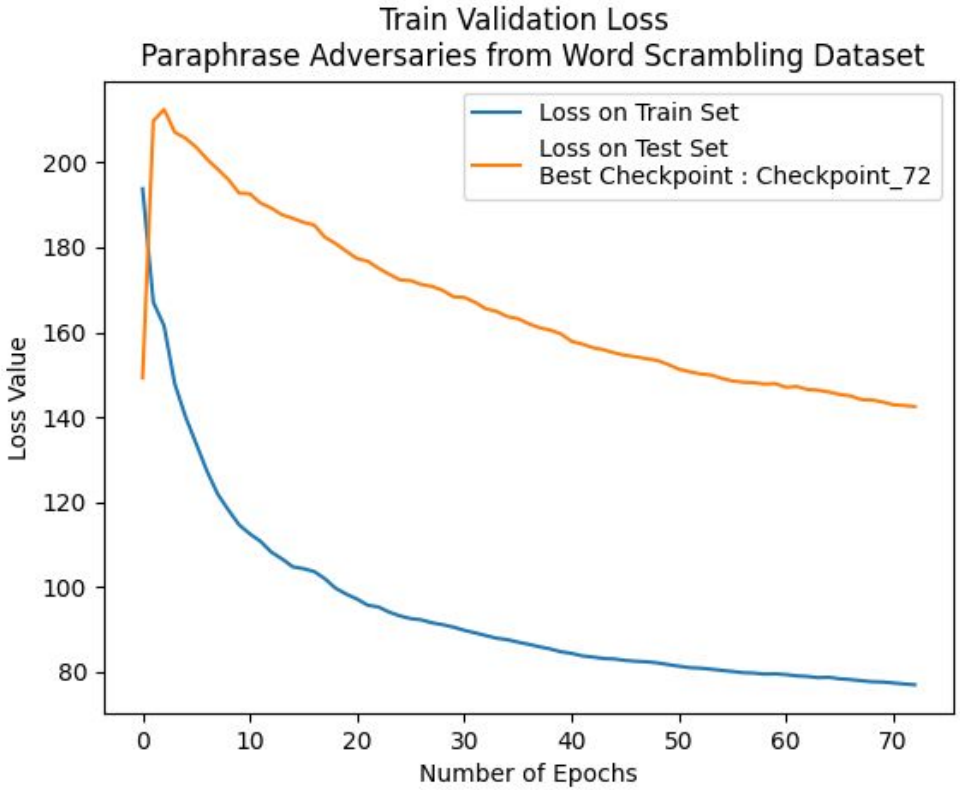
Results
&
Discussions

Quantitative Evaluation of PAWS Data

BLEU Score : 7.312 on test set of MSR Paraphrase dataset.

This score is from checkpoint : 72

Train & Valid loss graph : PAWS Dataset



Qualitative Evaluation of PAWS Data

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

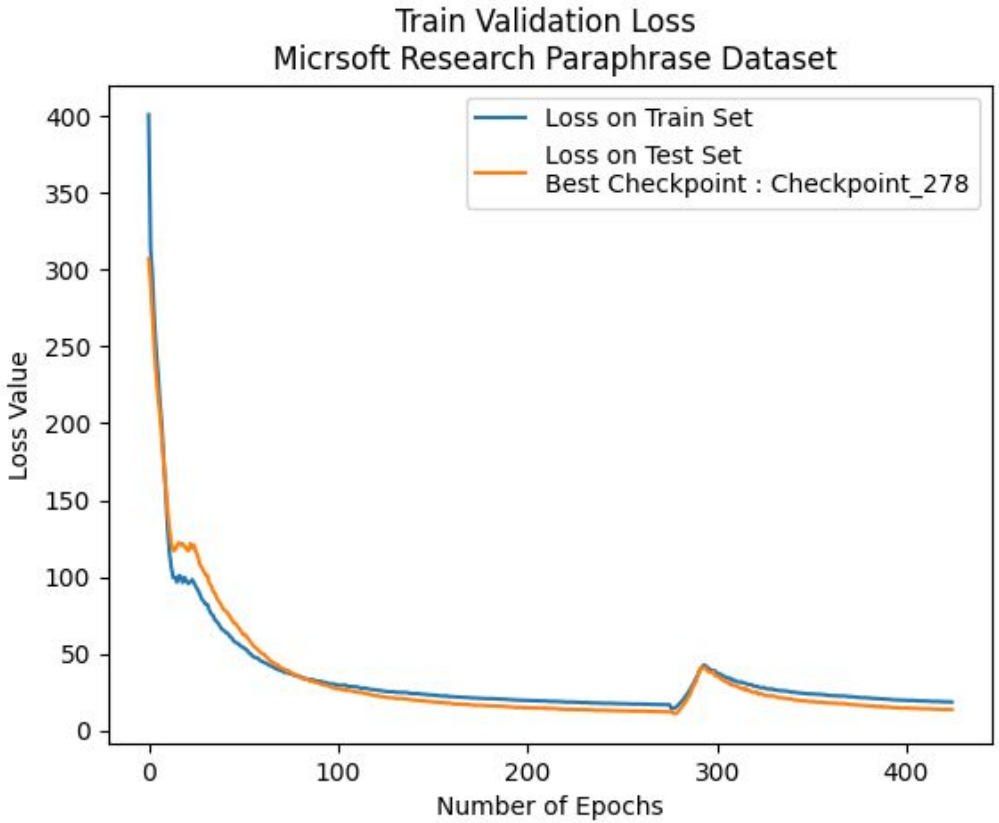
Predicted : THE FIRST FIVE ON THE AREA WERE THE FIRST AND LAST WITH THE FINAL

Quantitative Evaluation of MSR Data

BLEU Score : 53.439 on test set of MSR Paraphrase dataset.

This score is from checkpoint : 278

Train & Valid loss graph : PAWS Dataset



Qualitative Evaluation of MSR Data

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

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.

Thank You