Project Update 2: Text Summarization Using NLP

Mark Trovinger
Purdue Fort Wayne
tromv01@pfw.edu

Atharva Atre
Purdue Fort Wayne
atreaa01@pfw.edu

Navyaprabha Rajappa
Purdue Fort Wayne
rajan02@pfw.edu

1 Results

For the second round of updates, we focused on the more complicated and computationally expensive Transformer based models. Similarly to the first update, we used the CNN/Daily Mail dataset to fine-tune different Transformer based models.

1.1 Current

Model	ROUGE 1	ROUGE 2	ROUGE-L
BART	39.8079	18.036	24.8074
BERT	35.4348	15.5857	24.1659
Seq2Seq	0.099	0.0141	0.0942

The scores for BART and BERT were obtained from the entirety of the dataset, while Seq2Seq was trained on the first 500 rows.

1.1.1 BART

BART (Bidirectional and Auto-Regressive Transformers) is a denoising autoencoder pre-trained on corrupted documents that can be adapted to a number of different NLP tasks. BART can be thought of as a more generalized version of BERT, discussed below. BART was developed by Facebook AI in 2020, and has been fine trained on tasks such as text summarization, sequence classification, token classification, seqence generation, and machine translation.

1.1.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a powerful pre-trained natural language processing model that can be fine-tuned for various tasks, including text summarization. It uses a transformer architecture to process and understand natural language and has been shown to be highly effective in capturing the context and semantics of language.

We trained the last few layers of the BERT model on the CNN data set.

1.1.3 Seq2Seq

Seq2seq (Sequence-to-Sequence) is a type of neural network architecture that is commonly used for natural language processing tasks such as machine translation, text summarization, and speech recognition.It consists of two recurrent neural networks (RNNs) that work together to map an input sequence to an output sequence. The first RNN, called the encoder, reads the input sequence and generates a hidden state, which is then used by the second RNN, called the decoder, to generate the output sequence.

Like any other algorithm, there are advantages and disadvantages to seq2seq. Some of the advantages of are that Seq2Seq models can generate abstractive summaries, which means they can generate summaries that are not just a selection of sentences from the original document, but instead capture the main points and meaning of the document in a more concise manner. Also, Seq2Seq models are language-independent and can be trained on documents in multiple languages, making them highly adaptable for multilingual summarization tasks. In addition, Seq2Seq models can handle long documents better than TextRank, which can struggle with longer documents. Finally, Seq2Seq models can capture the context and meaning of the original document and generate summaries that are more faithful to the original meaning, whereas TextRank may miss some of the nuances and context.

The disadvantages include requiring large amounts of training data; Seq2Seq models require a large amount of training data to perform well, which can be a challenge for summarization tasks where large amounts of annotated data may not be available. Also, Seq2Seq models are more computationally expensive than TextRank, which can be a concern when processing large amounts of text. Unsurprisingly, Seq2Seq models can be more complex to implement than TextRank, which is a relatively simple algorithm.

1.2 Future

At this point, we will work towards further tuning the architecture that holds the most promise.

2 Analysis

As we can see from the results, both BART and BERT perform similarly, with BART holding a slight edge. We want to dive deeper into hyperparameter optimization to see which model performs

the best.

2.1 Future

Further analysis will likely focus on summarization tasks using different datasets. There are several datasets available that can be examined. Based on these scores, BART appears to be the best model for text summarization. It achieved the highest ROUGE-1 and ROUGE-L scores among the three models, indicating that it is able to produce more accurate summaries that preserve the meaning of the original text. BERT also performed relatively well, but its scores were slightly lower than BART. The seq2seq model, on the other hand, performed significantly worse than both BART and BERT, although it should be noted that it was trained on a much smaller subset of the dataset. Overall, BART seems to be the most effective model for text summarization based on these scores.

3 Problems

As was mentioned in the previous update, compute resources continued to be an issue, as the free Colab version was unable to train for more than 500 rows before timeout.

3.1 Proposed Solutions

The compute problem has been mitigated for the future due to the acquisition of a workstation containing the following hardware: Dual Xeon 6138 Gold CPUs, 128 GB of RAM, and an NVIDIA RTX A5000 GPU. This will allow us to train larger models ahead of the final presentation and paper.