Abstractive Summarization

AI 829: Natural Language Processing Mandate-3

Sarthak Harne
IMT2020032
IIIT Bangalore
sarthak.harne@iiitb.ac.in

***Abstract*—For the last mandate, I have made a couple more additional contributions. The contributions have been made to improving the quality of the dataset, improving the training method drastically and suggesting a metric to judge the summarization task better. The notebook with the contributions can be found here and the dataset used can be found here**

I. Introduction

The pipeline has changed significantly from Mandate-3, the pipeline used for the task can be seen in 1.

I have modified the EMNLP2017 Dataset [1] and added gold standard summaries for a small subset of the dataset to train the summariser on high quality data.

I used a summariser-discriminator combination to enrich the semantics of the model and its outputs as done in [2]. I have explained about this in more detail in the Training Section.

For evaluation, I compare the results of the last Mandate, and mention why rouge scores might not always be the best choice and show how Bert scores might better reflect the semantic similarity.

II. Dataset

As mentioned in Mandate-3, the dataset contained the news articles which the tweets linked to. But these articles where given in HTML format. To extract the article text from these HTML files required scraping scripts. Moreover, these articles all came from different sources, like NY Times, Washington Post, etc. so different scripts would be required to scrape different websites. Instead, I generated the summaries using ChatGPT. ChatGPT has an added advantage of also looking at opinions expressed by people in their tweets which is very crucial in a Tweet Abstractive Summariser. This would not be available in the scraped articles

The EMNLP2017 Dataset has one column for tweets and another for a paraphrase tweet which links to the same article. The tweets are generally different for each row, but the paraphrase are similar for a set of tweets. In other words, many tweets could be mapped to a single paraphrase. So, after dropping duplicates in the tweets columns, I grouped the dataset by the paraphrase, and added another column 'group_id'. The median group size was 12. So, for all groups with at least 12 tweets in them, I randomly selected, 12 tweets. Out of these, I chose 300 groups to generate a summary for due to time constraints.

To generate the summary, I first did basic preprocessing of the tweets in a group and then joining them using full-stops (.). Then I prompted ChatGPT to give me a summary of the tweets. I used this as a summary for the group these tweets belonged to. Using this, I created another small dataset which had group ids and their corresponding summaries, this is also available in the aforementioned dataset. This process can be automated via the ChatGPT API, but, unfortunately, it is proprietary.

Finally, for the train dataset, each row had

- the 'input' which is the input to the summariser, it is tokenized using the T5Tokenizer.
- the 'target' which is the label to the summariser, it is also tokenized using the T5Tokenizer.
- the 'input_text' which was a string of tweets corresponding to a group id joined using full-stops (.).
- the 'target_text' which is the basic preprocessed version of the summary generated by ChatGPT for that group id, this is useful to train the discriminator and also during evaluation.

The biggest improvement of using this dataset over the last mandate is presence of a high quality summary. Instead of just using a partition of a group as a summary for another partition, now we have a gold standard summary for the whole group.

III. Training Methodology

As mentioned in the last mandate, even though the Rouge scores were pretty high, that happened because of significant overfitting of the model. The model learnt to repeat input sentences and hence as the Rouge scores focus on measuring word similarity, we observe a high rouge score for the model.

At first, I tried to just train the model using the new, improved summaries, but as I was training the model, I observed that the Rouge scores were increasing continuously, but the Bert score (explained in the Evaluation part) kept

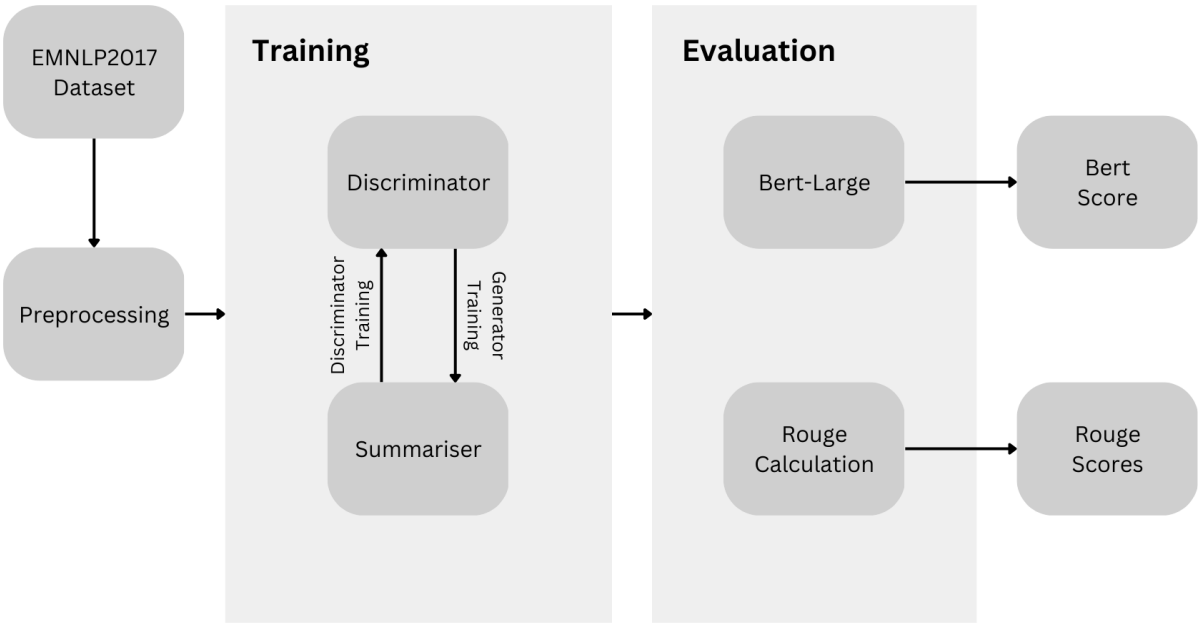

Fig. 1: Final Pipeline used

dropping. This again meant that the model was starting to learn the words and not the actual meaning. We will also see this further using some examples.

I realised that not enough information is being given to the model so that it can generate a summary similar to what a ”human” (here, ChatGPT) would do. An additional loss function was required to drive the summariser. For this, I used sn adversarial discriminator. The training methodology has been inspired from the training of GANs [2].

The discriminator I chose was Bert with a linear classifier layer on top, as it is known to understand semantics of language better (explained more in evaluation). The summariser model I used was T5.

Each iteration in the training loop (for a mini batch) is comprised of three steps:

- Training the discriminator to recognise the ”real” summary. The summary for that particular group is passed to the discriminator. We label the real summary as 1 to drive the Binary Cross Entropy Loss for the discriminator.
- Training the discriminator to recognise the generated summary. The input text for the group is passed to the summariser, where it generates a summary, this summary is passed to the discriminator after detaching from the compute graph for the summariser. We label the real summary as 0 to drive the Binary Cross Entropy Loss for the discriminator.
- Training the summariser via the sum of Language Modelling Loss and the Binary Cross Entropy Loss obtained after passing the generated summary through discriminator, with label 1.

The architecture can be seen in 2

I used separate optimisers (AdamW) for the summariser and the discriminator. During the third step of the training, only the summariser was updated.

The idea behind this was to train a model which could differentiate between what a gold standard summary and a generated summary. Further, training the summariser so that it can learn to ”fool” the discriminator and in the process, generate summaries close to gold standard summaries.

To check the effectiveness of using this methodology, I also trained the summariser, on the new dataset, without the help of the discriminator and have reported the findings. It becomes clear how the discriminator gives a major boost to both the Rouge and Bert scores.

It has been now shown multiple times that Bert is good at understanding semantics of language [3] [4]. Due to this a Bert based discriminator is very useful as it doesn't just look at the generated words, but also tries to look at the semantics of both the summaries to decide which is a generated and which is a real summary. We will exploit this fact during evaluation.

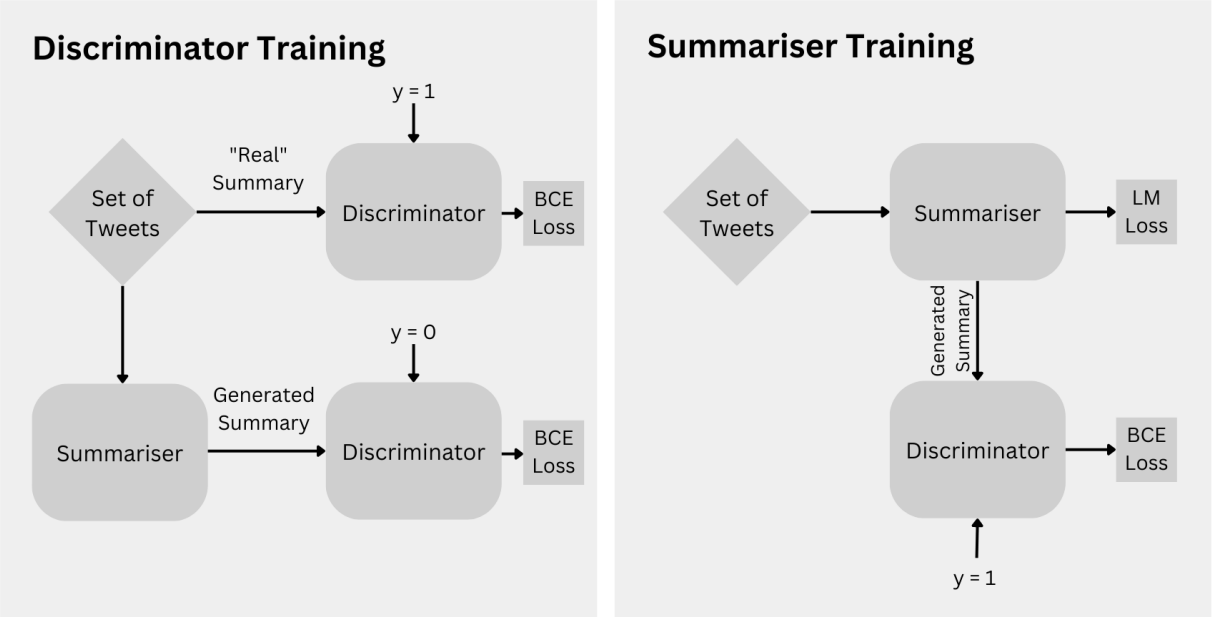

Fig. 2: Training Methodology

NOTE: For the basic preprocessing, I followed the same steps as in Mandate 2, which include:

- Expanding Contractions
- Removing Punctuations
- Converting to Lower Case

This step of preprocessing has been applied to both the tweet and the summary.

IV. Metrics

As mentioned in earlier Mandates, Rouge score might not be a suitable metric for abstractive summarization as it looks at the number of words and n-grams that co-occur in the target and the prediction.

As an evidence, we can see the following:

- Target: sammy lee the first asianamerican man to win olympic gold he still had to fight discrimination
- Prediction: he first asianamerican man to win olympic gold fought discrimin

This gives us a rouge scores of (57.7, 38.3, 53.3). High rouge scores indicate the commonality of words, but that also means that the prediction might just be a paraphrase of the tweets and not the summary.

After saying this, it is still important to take into account the commonality of words while judging the quality of summaries. But, we still need an additional metric to judge the semantic similarity.

As mentioned before, it has been shown that Bert is good at understanding semantics [3] [4], and hence a metric, Bert score, is devised. It works in the following manner:

- the contextual embeddings of each word in the two pieces of text are first generated using the BERT model
- these embeddings are then used to compute the cosine similarity between each pair of embeddings in the two pieces of text
- the BERT Score is the harmonic mean of the precision and recall of these cosine similarities
- precision = number of matching words in predicted and reference text / total number of words in predicted text
- recall = number of matching words in predicted and reference text / total number of words in reference text

On the surface, this might look similar to BLEU and Rouge scores, but here 'matching words' is defined to be words that have a cosine similarity score above a certain threshold. As Bert produces embeddings for the words, a cosine similarity based metric does not limit it to similarity of words, but focuses more on semantics.

V. Experiments, Evaluation and Results

To verify the usefulness of the adversarial training method, I first trained the model only on the new dataset, without

TABLE I: Scores for various experiments, scores multiplied by 100

Model Specs	Rouge-1	Rouge-2	Rouge-l	Bert Score
Without Discriminator	31.4	15.1	30.8	0.39
With Discriminator	43.6	21.1	32.3	0.60

the discriminator, see I. For which I received rouge scores of (31.4, 15.1, 30.8) and a Bert score of 0.39. But, when I used the adversarial training strategy, I received rouge scores of (43.6, 21.1, 32.3) and a Bert score of 0.60. This clearly showed the effectiveness of the training strategy.

Moreover, while training without the adversarial method, I observed that the rouge score went up gradually, while after rising for the first epoch, the Bert score started falling. On the other hand, the rouge scores and Bert score both went up with each epoch when using the adversarial strategy. This clearly showed how normal fine tuning drove the summariser to overfit while the adversarial strategy introduced semantics to the summaries. This will become very apparent after we look at some examples of the generated summaries. You can see the plots for the same in 3

For the sake of brevity, I have moved the examples to the Appendix section.

Moreover, we also observe that the discriminator has learned how to perfectly distinguish between a generated summary and a gold standard and predicts correctly with almost 100% accuracy. If this was not the case, and the discriminator did not perform well, we would not have a good loss which could drive the summariser to learn better.

Due to all of this, we can see that although not dispensable, rouge scores need to be supported by other metric which account for semantics, like Bert scores, especially for tasks like Abstractive Summarization.

NOTE: I used torchmetrics for calculation of both Rouge and Bert scores. Although a RoBERTa-large model is recommended to calculate the scores, I used a Bert-Large model due to computational constraints.

VI. Future Work

Some other improvements that can be made:

- Generate a bigger corpus of summaries, or, ideally, get human written summaries. This can also be done via scraping the news articles, but then the summaries won't account for the opinions expressed in the tweets.
- Auto Encoders can be trained and used before the cross-attention layers of T5 to introduce noise and avoid plain paraphrasing of the text.
- Better models like RoBERTa Large can be used to calculate the Bert scores to get better results.

References

[1] W. Lan, S. Qiu, H. He, and W. Xu, "A continuously growing dataset of sentential paraphrases," in *Proceedings of The 2017 Conference on Empirical Methods on Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2017, pp. 1235–1245. [Online]. Available: http://aclweb.org/anthology/D17-1127

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[4] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

VII. Appendix

Here are some example results:

- Input Text: 38 dead and 155 wounded in twin bomb attack near istanbul stadium who is really behind this atta. 29 dead and 166 wounded in twin bomb attack near istanbul stadium. terror in istanbul 29 dead 166 wounded in twin bomb attack in turkey. thoughts and prayers to the victims 38 dead and 155 wounded in twin bomb attack near istanbul stadium foxnews. 15 dead 69 wounded in twin bomb attack in istanbul. thoughts and prayers to the victims 38 dead and 155 wounded in twin bomb attack near istanbul stadium foxnews. 38 dead and 155 wounded in twin bomb attack near istanbul stadium what goes around comes around. 38 dead and 155 wounded in twin bomb attack near istanbul stadium what goes around comes around
- Gold Standard: a twin bomb attack near an istanbul stadium has resulted in 38 dead and 155 wounded the attack has been widely condemned and thoughts and prayers are with the victims there are no clear indications yet as to who is responsible for the attack
- Summariser without Discriminator: turkey: 38 dead, 155 wounded in twin bomb attack near istanbul stadium. foxnews: 166 wounded in twin bomb attack near istanbul stadium. foxnews: russian prime minister vladimir putin vladimir putin vladimir putin.
- Summariser with Discriminator: the paragraphs report on a twin bomb attack near an istanbul stadium that left 38 people dead and 155 wounded the focus is on the attack and the potential for further violence in turkey with coverage highlighting the reporting by fox news and other sources
- Input Text: at least 27 people killed and 57 others wounded when blasts went off near car spare parts shops in al sinak. two blasts in central baghdad kill at least 28. double bomb attack hits al sinak market at least 21 killed and dozens wounded as blasts triggered by. at least 18 killed as bomb attacks rock central baghdad ajenglish. at least 21 killed and dozens wounded as blasts triggered by suicide bomber and explosive ajenglish. 2 blasts in central baghdad kill at least 18 police say afp al jazeera. dozens also injured as two bombs explode at busy market in central al sinak neighbourhood. at least 21 killed and dozens wounded in baghdad after a suicide bombing

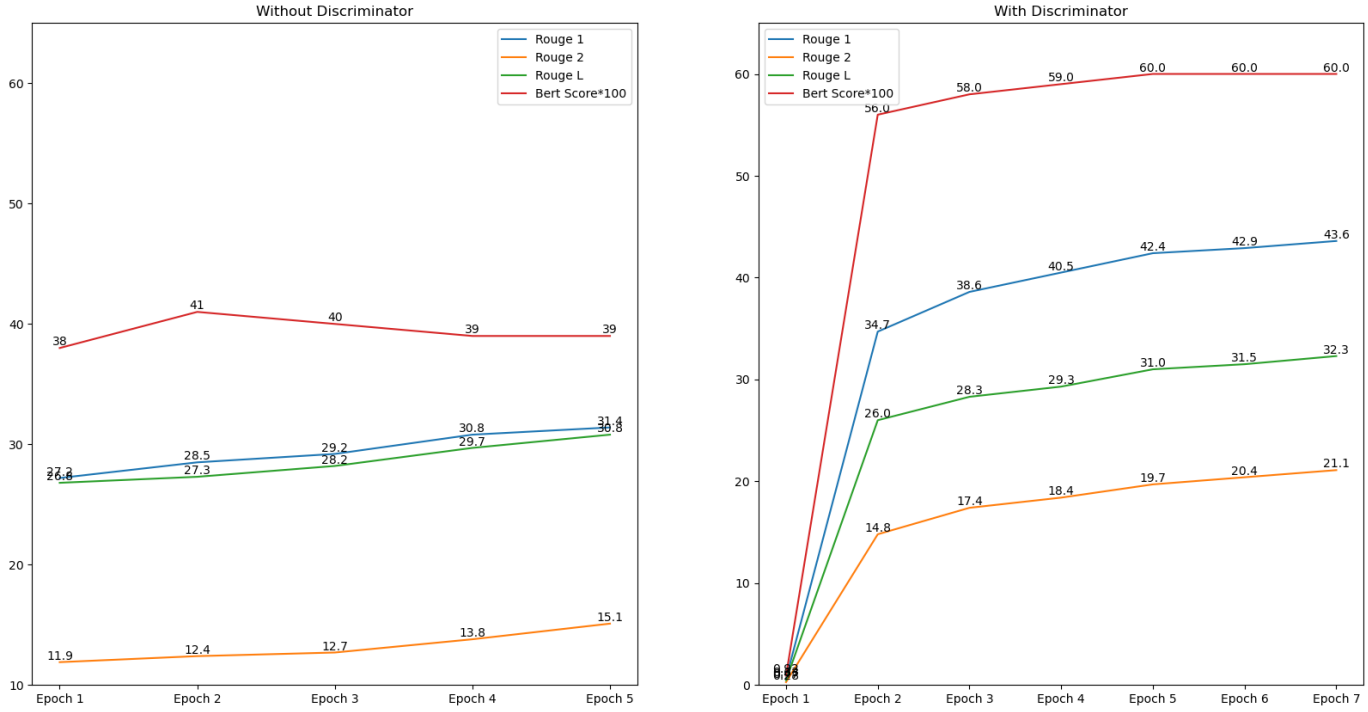

Fig. 3: Scores over the epochs

- Gold Standard: two blasts in central baghdad specifically in the al sinak market resulted in the deaths of at least 27 people and wounded 57 others the explosions occurred near car spare parts shops and were triggered by a suicide bomber and an explosive the attack took place in a crowded market causing dozens of injuries the death toll was later updated to at least 28 people the blasts were part of a series of bombings in central baghdad resulting in the deaths of at least 18 people
- Summariser without Discriminator: double bomb attack hits al sinak market killing at least 28 people. explosions triggered by suicide bomber and explosive ajenglish. at least 21 killed and dozens wounded in baghdad after a suicide bombing.
- Summariser with Discriminator: two bombs explode in central baghdad near car spare parts shops killing at least 28 people and injuring 57 others the explosions occurred near a busy market in al sinak and were triggered by a suicide bomber and explosive the blasts occurred in the central market and were triggered by the suicide bomber and explosive ajenglish the blasts killed at least 21 people and injured dozens more
- Input Text: the new york times how to pounce on best credit card offers. how to pounce on best credit card offers before banks pull them. how to search for the best credit card deals before the disappear. how to pounce on best credit card offers last chance for freemoney just pick it up. 1500 in your pocket ronlieber explains how to pounce on the best credit card offers. how to pounce on best credit card offers por ron lieber en nyt. how to pounce on best credit card offers by ron lieber the new york times. new york times most viewed stories how to pounce on best credit card offers
- Gold Standard: the article discusses how to take advantage of credit card offers before they expire and how to search for the best deals ron lieber explains how to earn free money by pouncing on credit card offers and the article is one of the most viewed stories on the new york times website
- Summariser without Discriminator: the new york times explains how to pounce on the best credit card offers. ron lieber explains how to search for the best credit card deals before they disappear. lieber: i'm a pouncer, and i'm not a pouncer.
- Summariser with Discriminator: the new york times has compiled a list of tips on how to pounce on credit card offers before they disappear the list includes tips on how to search for the best credit card deals before banks pull them some articles have been viewed by thousands of people including ron lieber some articles have been criticized for not pinning on the best credit card offers while others have criticized the lack of transparency in the credit card industry
- Input Text: excellent executive decision trump wants to cancel air force one order from boeing. build it in mexico it will be cheaper that way trump wants to cancel air force one order from boeing. when did republicans become so fond of regulation trump wants 2cancel air force one order from boeing. trump lies again says jet costs more than 4 billion amp wants to cancel air force one order

from boeing. trump lies again says jet costs more than 4 billion amp wants to cancel air force one order from boeing. presidentelect realdonaldtrump wants to cancel a government order for a new air force one. amazing leadership and not even sworn in trump wants to cancel air force one order from boeing. trump wants to cancel air force one order from boeing what about the money for security at trump tower

- Gold Standard: donald trump wants to cancel the order for a new air force one from boeing claiming it is too expensive over 4 billion and suggesting building it in mexico would be cheaper he also suggests using his own jet and charging taxpayers there are criticisms of his decision including concerns about security at trump tower and the accuracy of his cost estimates
- Summariser without Discriminator: julian zelizer: trump lies again says jet costs more than 4 billion amp wants to cancel order. zelizer: amazing leadership and not even sworn in trump wants to cancel order from boeing. he says trump lies again says jet costs more than 4 billion amp wants to cancel order. zelizer: trump lies again says jet costs more than 4 billion amp wants to cancel order.
- Summariser with Discriminator: the paragraphs report on presidentelect trumps decision to cancel an order for a new air force one from boeing which is a good executive decision and a good example of the importance of executive discretion in the government some articles mention the lack of funding for security at trump towers and the lack of transparency in the government the tone is critical of trumps decision and the potential impact on the economy