Text Summarization using BART on xsum and samsum datasets

Vazgen Tadevosyan
ENGL.681 Natural Language Processing I
Rochester Institute of Technology
vt7313@rit.edu

I. Introduction

Recent transformer-based language models provide impressive results on various tasks ranging from machine translation to question answering. Large pre-trained models such as BERT and GPT-2 can easily be finetuned for different downstream tasks. In this work, I analyze BART and PEGASUS models' ability to summarize text and Dialogue. The code can be found here

https://github.com/paligonshik/Dialogue_Text_Summarization

II. Setup

Hugging Face is an AI platform created by the AI community. Pretrained models can save the time and resources needed to train a model from scratch while lowering computing costs. Recent pre-trained models may be downloaded and trained using the APIs and tools provided by Transformers in Hugging face. After training, models can be deployed in hugging face using the user's access token and login credentials. I used the pytorch-nightly framework to utilize M1 GPU capabilities. Another reason why I decided to use the platform for this project is because they have everything I need in one place: the model, the data, and the libraries. I also learned how to do the deployment.

III. Datasets

Narayan et al. (2018)Extreme Summarization **XSum** dataset is a dataset for evaluating abstractive single-document summarizing. The features of the dataset are documents with their respective summarized text. It consists of 226,711 news stories and a one-sentence summary, The articles present a wide range of topics from BBC journals published between 2010 and 2017. The topics are News, Politics, Sports, Weather, Business, Technology, Science, Health, Family, Education, Entertainment, and Arts.

In Figures 1-3 you can see some insights of the data.

Gliwa et al. (2019)**SAMSum** dataset contains about 16k facebook messenger talks. Linguists with English fluency created and recorded conversations. Conversations can take on a variety of styles and topics, including casual, semi-formal, and formal ones, and they may also use slang terms, emoticons, and typos. It also contains summaries as annotations to the dialogues.

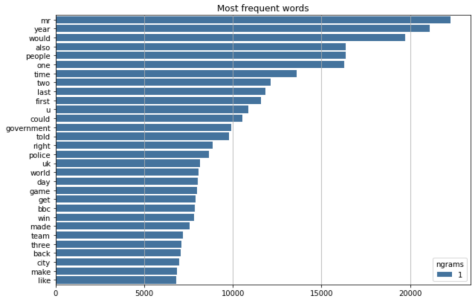

Figure 1. Most Frequent words in document after initial preprocessing (xsum)]

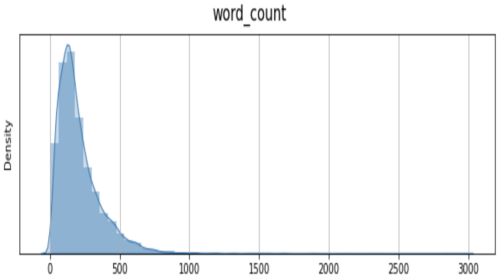

Figure 2. Distribution of the number of words in the document (xsum)]

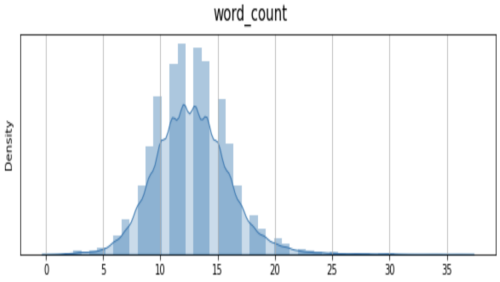

Figure 3. Distribution of the number of words in the summary (xsum)]

IV. Text Summarization

There are two different types of text summary techniques extractive summarization and abstractive summarization. **Abstractive Summarization** chooses a subset of sentences from the text to create a summary; it identifies the critical points in the text and inserts new words or phrases. it requires a text generation component, such as a decoder. Since **Extractive Summarization** techniques are often relatively straightforward, choosing which sentences to include in the summary is framed as a binary classification task for each phrase in the text. Therefore, accuracy may be used to measure extractive summarization performance. However, in summarization tasks, scientists use **ROUGE** Score to evaluate the model's performance. A measure created specifically for automatic text summarization is called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Lin (2004). Most articles evaluate their approaches and models using ROUGE-N (ROUGE-1, ROUGE-2) and ROUGE-L. ROUGE-N is a recall-inspired measure that is determined by dividing the total number of N-grams in the reference summary by the number of N-grams shared by the reference (human-written) summary and the AI-written summary.ROUGE-1 thus only focuses on the unigrams in the summary, whereas ROUGE-2 only targets the bi-grams.

V. Models

Most NLP models, such as **BERT** Devlin et al. (2018) or **GPT-2** Radford et al. (2018), are too large to be trained for this particular need. Even finetuning them needs some powerful processing. I found out that Facebook and Google had already solved the problem. Facebook researchers designed a new **BART** Lewis et al. (2020) model, which has an autoregressive decoder; it can be directly finetuned for sequence generation tasks such as abstractive summarization. It can be a generalization of BERT and GPT2 as it includes bidirectional networks too.

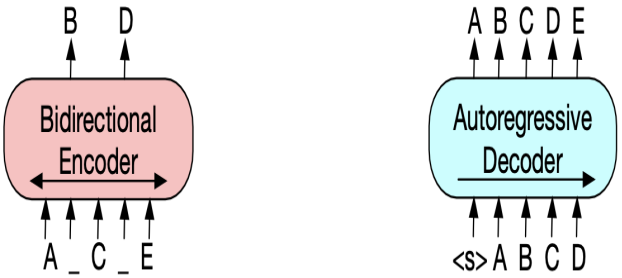

Figure 4. Token Masking in BERT VS GPT2

See Fig. 4, BERT randomly masks tokens s, and the document is encoded bidirectionally. The prediction of masked tokens is independent, thus preventing BERT from generating new sentences. GPT autoregression to predict the masked tokens, meaning GPT can be used for generation. However, newly taken depend only on leftward context, so it cannot learn bidirectional dependencies. Meanwhile, BART (Fig. 5) uses

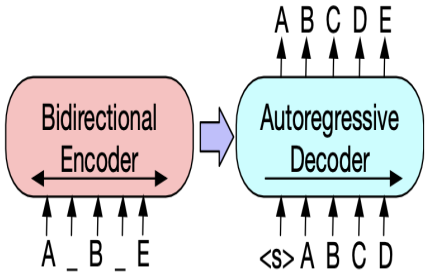

Figure 5. Token Masking by BART

both components. Arbitrary noise masking is possible since the encoder's inputs and outputs do not have to match. Here, portions of the text have been changed to disguise symbols to distort the text. An autoregressive decoder is used to determine the probability of the original document (right) after the damaged document (left) has been encoded using a bidirectional model. An uncorrupted document is sent to the encoder and decoder for fine-tuning, and representations from the decoder's final hidden state are passed for the prediction. Pre-training with Extracted Gap-Sentences for Abstractive Summarization (**PEGASUS**) Zhang et al. (2019) was created by Google AI in 2020. They suggest pre-training large encoder-decoder models based on Transformer on enormous text corpora with a new self-supervised goal. Unlike random word masking in BERT, several sentences from a document are masked in PEGASUS. These sentences can be predicted using the same autoregressive and bidirectional approach. A document with missing sentences is the input; PEGASUS will find them; the output is composed of the recovered missing sentences. To create a masked language model, PEGASUS employs a pre-trained encoder.

VI. Results

After initial preprocessing, the sentences and words are segmented into tokens; then, a vocabulary is constituted to encode text into tokens and vice-versa. I used facebook-bart-base model for initializing parameters. Then it was finetuned on my datasets. The parameters are updated through the minimization of cost functions through updating strategies to update the parameters defined by the optimizers. I used weight decay for regularization and AdamW for optimising learning process. On average, it took more than 6 hours of training, even on a small sample of the data. Here I compare evaluation of BART and Pegasus after three epochs.

It is seen from Figs 6-7 that, on average, Pegasus provides higher scores than traditional BART. ROUGE 1, 2 refer to the size of n-grams used to calculate the score. Other metrics are some further extensions of multiple n-gram pairs averaging their results.

A. Examples

In Fig. 8, the summary's prediction is done using the Extractive TextRank approach. We can think of as baseline,

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	RougeI	RougeIsum	Gen Len
1	1.721400	0.458949	33.337800	11.761600	26.467900	26.410800	19.370000
2	0.466500	0.450347	33.675200	12.511400	27.170300	27.160600	19.715000
3	0.400800	0.453694	34.173200	12.750700	27.659500	27.656700	19.745000

Figure 6. Evaluation of BART

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	RougeI	RougeIsum	Gen Len
1	0.541900	0.358164	49.186900	24.529500	39.786200	39.768500	23.965000
2	0.216300	0.367485	50.832700	24.929000	40.741500	40.661100	26.410000
3	0.135200	0.415083	51.023700	24.166800	39.646000	39.612500	29.625000

Figure 7. Evaluation of PEGASUS

it simply extracts the most important sentences. I highlight them to show which sentences and words are extracted.

Full Text

A fire alarm went off at the Holiday Inn in Hope Street at about 04:20 BST on Saturday and guests were asked to leave the hotel. As they gathered outside they saw the two buses, parked side-by-side in the car park, engulfed by flames. One of the tour groups is from Germany, the other from China and Taiwan. It was their first night in Northern Ireland. The driver of one of the buses said many of the passengers had left personal belongings on board and these had been destroyed. Both groups have organised replacement coaches and will begin their tour of the north coast later than they had planned. Police have appealed for information about the attack. Insp David Gibson said: "It appears as though the fire started under one of the buses before spreading to the second. "While the exact cause is still under investigation, it is thought that the fire was started deliberately."

Predicted Summary

A fire alarm went off at the Holiday Inn in Hope Street at about 04:20 BST on Saturday and guests were asked to leave the hotel. One of the tour groups is from Germany, the other from China and Taiwan. It was their first night in Northern Ireland. Police have appealed for information about the attack. Insp David Gibson said: "It appears as though the fire started under one of the buses before spreading to the second.

Figure 8. Extractive TextRank example

Now let's look at Fig. 9 and how BART generates new sentences using only the most important words from the document.

Last but not least, in last figure I represent the output of PEGASUS on a sample dialogue.

Wanda: Let's make a party! Gina: Why? Wanda: beacuse. I want some fun! Gina: ok, what do u need? Wanda: 1st I need too make a list Gina: noted and then? Wanda: well, could u take yours father car and go do groceries with me? Gina: don't know if he'll agree Wanda: I know, but u can ask :) Gina: I'll try but theres no promisess Wanda: I know, u r the best! Gina: When u wanna go Wanda: Friday? Gina: ok, I'll ask

Prediction:*Wanda wants to have a party. Wanda and Gina will go shopping on Friday. Gina will take her father's car and go shopping with Wanda.*

Full Text

A fire alarm went off at the Holiday Inn in Hope Street at about 04:20 BST on Saturday and guests were asked to leave the hotel. As they gathered outside they saw the two buses, parked side-by-side in the car park, engulfed by flames. One of the tour groups is from Germany, the other from China and Taiwan. It was their first night in Northern Ireland. The driver of one of the buses said many of the passengers had left personal belongings on board and these had been destroyed. Both groups have organised replacement coaches and will begin their tour of the north coast later than they had planned. Police have appealed for information about the attack. Insp David Gibson said: "It appears as though the fire started under one of the buses before spreading to the second. "While the exact cause is still under investigation, it is thought that the fire was started deliberately."

Predicted Summary

Two tour groups have been attacked by two buses in Belfast.

Figure 9. Abstractive BART example

The deployed model can be found here https://huggingface.co/Paligonshik/pegasus-large-finetune-xsum

VII. Conclusion

Transformer-based models BART and Pegasus perform reasonably well on text and dialogue summarization. Even now, they can serve as useful plugin applications in many domains, from zoom meeting discussions to researching papers. However, there are few studies on dialogue summarization where there are more people than two. In the future, I plan to analyze those models' capability to deal with multiperson dialogues. Several generated sentences, in summary, is another area to do research. Sometimes one sentence is not enough to address critical points of the text and dialogue; on the other hand, very long summaries won't make any sense for their purpose. Thus finding an optimal number of sentences to generate a summary is a very necessary and interesting area to do further research.

References

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL http://arxiv.org/abs/1810.04805. cite arxiv:1810.04805Comment: 13 pages.

B. Gliwa, I. Mochol, M. Biesek, and A. Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019.