COSE474-2024F: Final Project
BART-based Article Summarization Model

Park Subeen

Abstract

In today's fast-paced world, people often lack the time to read full articles. Simultaneously, large-scale pre-trained language models have shown remarkable performance in various natural language processing tasks, making them ideal for providing concise and meaningful article summaries. This project investigates automatic article summarization using BART, a widely used model for text summarization, applied to the CNN/Daily Mail dataset. The performance of BART is compared with that of the state-of-the-art model, Pegasus. Experimental results indicate that while BART shows a more balanced reduction in loss and better performance in ROUGE scores, Pegasus demonstrates a faster inference speed with a lower evaluation loss. These findings highlight the strengths and limitations of both models in article summarization tasks, providing insights into their potential applications and areas for further improvement.

1. Introduction

These days, people usually don't read articles because of their busyness. At the same time, large-scale pre-trained language models have demonstrated powerful performance across various natural language processing tasks. Therefore, language models are suitable for providing meaningful article summaries to these people.

This project aims to explore automatic article summarization with BART, which is the widely used models in the test summarization task (Ahn & Park, 2022), using the CNN/Daily Mail dataset. Additionally, its performance will be compared with other state-of-the-art (SOTA) model, Pegasus, and we will explore the strength and limitation of BART in article summaries.

2. Problem Definition & Challenges

Text summarization is the task of generating short and concise summaries from long documents while preserving key information. The challenge in article summarization is to maintain the core content without distorting the fact delivered by the original article. Overfitting and bias of the model could also be the challenges.

Therefore, this project will utilize proper preprocessing and dropout.

3. Related Works

Initially, statistical model was usually used to generate text summaries. But recently, research on summarization using artificial neural network models has been more conducted. The developments especially aim to improve the model's ability of apprehending textual content within pre-learning as well as additional pre-training steps. Since traditional sentence-level masking has limitations in focusing on repetition rather than meaning, there is also research on masking focusing on the meaning of sentences. (Jeon, 2024)

Although there is also a study on summarizing Korean articles, it found that it was difficult to make an accurate and core summary in Korean. (Seol, 2018)

4. Datasets

The CNN/Daily Mail dataset from Hugging Face Datasets will be used for training and evaluating the model. It contains over 300,000 news articles along with human-written summaries, offering diverse categories and varying levels of complexity are also diverse. Therefore, it is well-suited dataset for evaluating the performance of BART on article summarization tasks.

5. State-of-the-art Methods and Baselines

We will compare the performance of BART with that of Pegasus. Metrics such as ROUGE will be used to assess the quality of the generated summaries in terms of precision and recall. (OpenAI, 2024)

The baseline of the loss value at first epoch is 2.5, and those of ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L) are 40%, 15%, and 30%, respectively.

6. Experiment

6.1. Setup

Table 1. Computing Resources

RESOURCE	VERSION
CPU	X86_64
OS	LINUX 6.1.85+
PYTORCH	2.5.1+CU121
CUDA	12.1

Table 1 shows the computing resources of this project.

Algorithm 1 Text Summarization with BART

Input: Dataset D, Pretrained BART Model M, Tokenizer T, Hyperparameters H
Output: Trained Summarization Model S
Split dataset D into training set D_{train} and validation set D_{val}.
Initialize pretrained BART model M and tokenizer T.
Configure hyperparameters H (e.g., learning rate, batch size, epochs, etc.).
for each epoch e in E **do**
  for each batch (x, y) in D_{train} **do**
    Tokenize input text x and target summary y using tokenizer T.
    Create tokenized input input_ids and attention masks for x.
    Create tokenized labels labels for y (target summary).
    Feed input_ids, attention masks, and labels into model M.
    Compute loss $\mathcal{L}$ (e.g., cross-entropy loss) using model predictions and labels.
    Backpropagate $\mathcal{L}$ and update model parameters using optimizer.
  end for
  Evaluate model M on D_{val} using ROUGE metrics.
  Save model parameters if the validation performance improves.
end for
return Trained Summarization Model S

Algorithm 1 shows the pseudocode of this project.

This project will use 'facebook/bart-base' for a model. It also will be conducted with a data ratio of 80% for training, 10% for validation, and 10% for testing, a batch size of 4 and 2 epochs. In addition, ROUGE metrics will be used for the evaluation of quantitative results. Furthermore, the performance of the proposed model will be compared with Pegasus.

6.2. Preprocessing

The 'preprocess_function' prepares the input articles and corresponding reference summaries for training. First, it adds the prefix "Summarize: " to each article in the dataset to specify that the task is summarization. The articles are then tokenized using the tokenizer, with a maximum length of 512 tokens, and padding and truncation are applied to ensure consistent input lengths.

For the reference summaries, the tokenizer is applied to the highlights field (the summaries) with a maximum length of 128 tokens. The padding and truncation are also applied to the summary tokens.

The labels (i.e., the reference summaries) are processed such that any padding tokens are replaced with -100. This step is necessary because padding tokens should be ignored during the model's loss calculation. Finally, the processed input tokens, attention masks, and summary tokens are returned for training.

6.3. Results

6.3.1. QUANTITATIVE RESULTS

Table 2. Loss Results of BART

EPOCH	TRAINING LOSS	VALIDATION LOSS
1	2.7286	2.2547
2	2.1814	2.2187

Table 2 shows the training loss and validation loss for each epoch of the BART model. The loss value decreases in both the training and verification datasets as the learning progresses.

Table 3. Evaluation Results of BART

METRIC	VALUE
EVALUATION LOSS	2.1324
EVALUATION RUNTIME (S)	160.3784
SAMPLES PER SECOND	0.624
STEPS PER SECOND	0.156

The evaluation results of the BART model is like Table 3. An Evaluation Loss (2.1324) lower than Validation Loss (2.2187) suggests that the model has a certain level of generalization ability on training and evaluation data. However, lower processing speeds will need to be improved later with better execution environments.

Table 4. ROUGE Results of BART

METRIC	SCORE
ROUGE-1	40.53%
ROUGE-2	19.24%
ROUGE-L	28.69%
ROUGE-LSUM	38.64%

The performance of the BART model on the test set was also evaluated using ROUGE metrics, and the results are as Table 4.

ROUGE-1 indicates that the key words are highly consistent between the generated summaries and the actual summaries. ROUGE-2 reflects the limitation of model in capturing multi-word patterns. ROUGE-L suggests that the model preserves the structural similarity at a decent level. Finally, ROUGE-Lsum reflects the overall sequence-level similarity.

Although this model achieved good performance in ROUGE-1, there are areas that need to be improved for other metrics, especially fluency and consistency. In conclusion, it can be seen that numerical results similar to baseline were obtained overall.

6.3.2. COMPARISON WITH PEGASUS

The Pegasus-xsum model is used for the comparison.

Table 5. Loss Results of Pegasus

EPOCH	TRAINING LOSS	VALIDATION LOSS
1	2.4036	1.9694
2	2.0853	1.9469

The training and validation losses of the Pegasus model are lower than those of the BART model. However, the rate at which the loss value decreases per epoch is higher for the BART model, with the training loss decreasing by about 0.55 and the validation loss by 0.04. In contrast, the Pegasus model shows a smaller reduction, with the training loss decreasing by about 0.32 and the validation loss by 0.02. This suggests that while the Pegasus model fits the training data more closely, the BART model has a better learning progression.

Table 6. Evaluation Results of Pegasus

METRIC	VALUE
EVALUATION LOSS	1.9108
EVALUATION RUNTIME (S)	4.1103
SAMPLES PER SECOND	24.329
STEPS PER SECOND	6.082

The evaluation loss of the Pegasus model is also lower than that of the BART model. Additionally, the Pegasus model demonstrates significantly faster runtime under the same experimental environment.

Table 7. ROUGE Results of Pegasus

METRIC	SCORE
ROUGE-1	31.42%
ROUGE-2	14.17%
ROUGE-L	21.47%
ROUGE-LSUM	21.51%

Despite achieving lower loss values, Table 7 illustrates that the ROUGE scores of the Pegasus model are lower than those of the BART model. This suggests that the Pegasus model might be overfitting the training data, leading to lower generalization perforance. Consequently, the quality of summaries produced by the Pegasus model may be inferior to those generated by the BART model.

In conclusion, this comparison indicates that the BART model achieves better overall performance on the article summarization task. Considering the balance between loss values and ROUGE scores, the BART model demonstrates more consistent and reliable results.

6.3.3. QUALITATIVE RESULTS

Based on the comparison between arbitrary original highlights and generated summaries (Appendix A), it can be said that the BART model summarizes important information well in natural sentences while removing unnecessary parts from the original article. However, some important details may be omitted or the facts may change somewhat.

Overall, the BART model can be said to have very effective results in text summarization.

7. Future Direction

Based on our findings, future research can focus on balancing accuracy, learning time, and inference speed and applying various methods, such as model pruning and adding other evaluation metrics, to optimize the performance of text summarization models such as BART and Pegasus. It can also promote performance improvements in special areas such as medical or legal texts as well as articles. Developing multilingual summary models will also be a major challenge for future research. These advances will ensure that text summarization systems provide efficiency in different contexts.

References

Ahn, Y.-P. and Park, H.-J. Fine-tuning of attention-based bart model for text summarization. Technical report, Korea Institute of information and Comminication Engineering, 2022.

Jeon, M. Deep learning-based natural language summary model learning methodology. Technical report, Graduate school of Business IT, Kookmin University, 2024.

OpenAI. Chatgpt, 2024. URL https://openai.com/chatgpt. Accessed: 2024-11-29.

Seol, K. Design and implementation of automatic summary system of korean news using deep learning models. Technical report, Graduate school of engineering, Hanyang University, 2018.

A. Summary Examples

Original Highlight	Generated Summary
Aung San Suu Kyi is released Saturday. She has been under house arrest for much of the past two decades. She has defiantly challenged the authority of the military junta. She likens Myanmar's plight to South African apartheid.	Aung San Suu Kyi is the very embodiment of Myanmar's long struggle for democracy. The 65-year-old human rights activist has endured house arrest for much of the past two decades and, perhaps, has become the world's most recognizable political prisoner.
An energized conservative electorate helps Republicans to historic gain in midterms. Republicans nab at least 60 more House seats, based on CNN analysis of exit poll data. Senate Majority Leader Harry Reid defeats Republican Sharron Angle in Nevada. President Obama calls House Minority Leader John Boehner to congratulate him.	President Barack Obama called House Minority Leader John Boehner of Ohio to congratulate him. The two discussed working together to focus on the top priorities of the American people, which Boehner has identified as creating jobs and cutting spending.
Thomas Craig, 91, drove through a red light and knocked over elderly lady. Brenda Forster, 89, was killed as she walked over pedestrian crossing. Retired taxi driver pleaded guilty and was handed suspended jail sentence. Miss Forster's daughter said she was angry that he was still driving.	Brenda Forster, 89, was hit by Thomas Craig's car at a pedestrian crossing after he failed to notice a red light. The retired taxi driver was handed a suspended jail sentence after pleading guilty to causing death by dangerous driving in June last year.
Manchester City beat Chelsea 2-0 in the English Premier League. Goals from Yaya Toure and Carlos Tevez seal win at Etihad Stadium. City 12 points behind local rivals Manchester United in EPL title race. Swansea thrash minnows Bradford to win the English League Cup.	Manchester City beat European champions Chelsea 2-0 in the Premier League on Sunday. Newcastle pulled six points clear of the relegation zone with an emphatic 4-2 win over fellow strugglers Southampton.
72 bodies found floating in a river or in three mass graves inside Rifles' compound. Fifty of the dead were confirmed to be army officers. Standoff started Wednesday when Rifles troops rebelled against commanders. More than 160 were inside Bangladesh Rifles headquarters when mutiny erupted.	More than 70 army officers still missing – and presumed killed – after a deadly uprising by paramilitary forces last week. The bodies were found floating in a river or in three mass graves. Prime Minister Sheikh Hasina is trying to appease an army that demands the killers be punished.