Boosting Query-Based Summarization by Exploiting Query Relations

Hailiang Dong, Yangxiao Lu, Priyanshi Shah, Adit Ketan Shah

Motivation and Introduction

- Meetings remain the go-to tool for collaboration, with 11 million meetings taking place each day in the USA and employees spending six hours a week, on average, in meetings. The result is summaries that capture the essence of a meeting and allow attendees to quickly catch up. It is difficult to condense or put together a brief summary that includes all the important details. Alternatively, summarization systems should adopt a more flexible and interactive approach that allows people to express their **interests** and **caters** to their diverse intents when generating summaries.

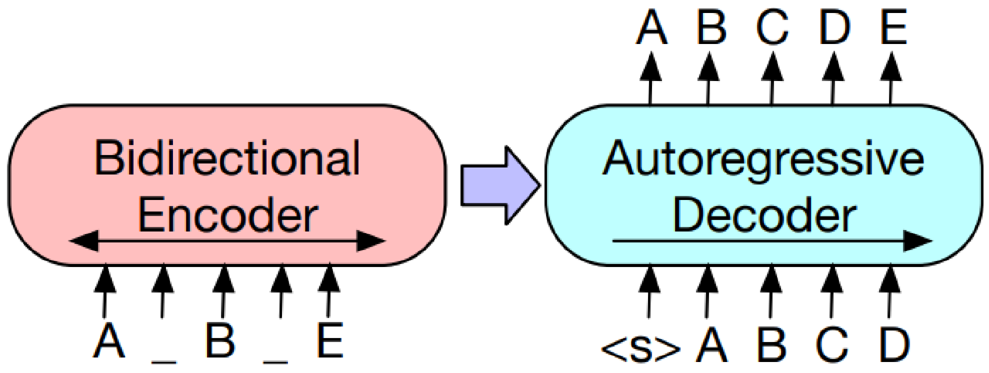
- We aim to leverage the power of fine-tuned **Facebook's BART model** to generate high-quality query-based summaries for meetings with their query histories, which can be beneficial for users who want to quickly get an idea of the meetings.

- Our project will involve **data preprocessing**, **fine-tuning** the BART model with user query histories, **generating** summaries for meetings, and finally **evaluating** the model performance through ROUGE scores.

- The dataset we are planning to use for this task is the QMSum dataset.

BART Model

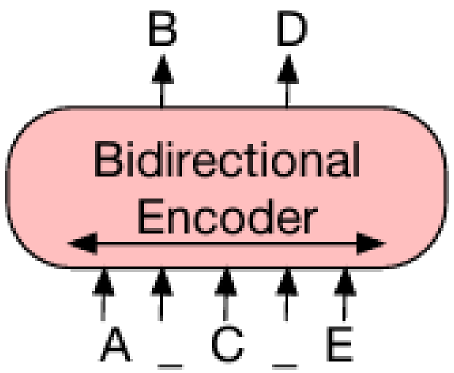
BART (Bidirectional and Auto-Regressive Transformers) model is a **pre-trained** transformer-based language model that is designed for **sequence-to-sequence** tasks.
Like BERT, this model can be applied to downstream tasks with limited data setting through fine-tuning.

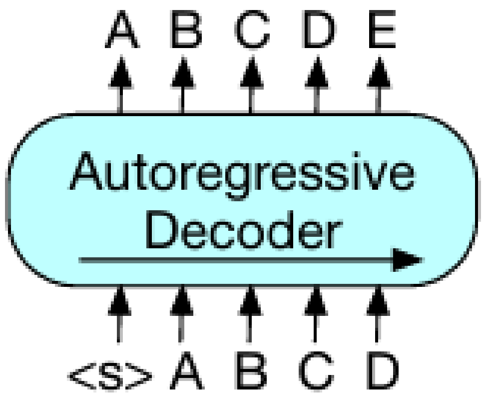


Key differences to BERT and GPT models:

- Designed to perform sequence to sequence (both auto-encoding and auto-regression) tasks
- Can generate output sequences using both left-to-right and right-to-left contexts
- Uses a **denoising** autoencoder training objective, while BERT and GPT models use masked language modeling

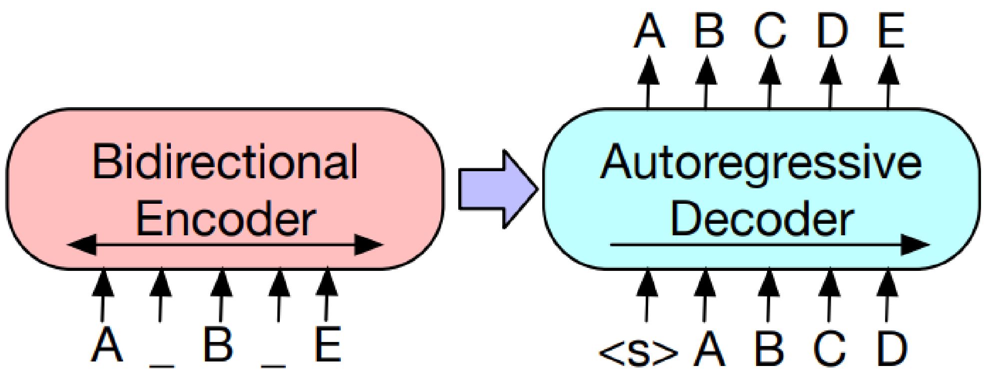
BART Model





BERT: Random tokens are replaced with masks

GPT: Tokens are predicted auto-regressively.
Then GPT can work for text generation.



BART: Masked input texts with encoder + autoregressive decoder.
Inputs need not be aligned with decoder outputs, allowing arbitrary noise transformations

Figures from BART, Lewis, Mike, et al.

Our Improvement

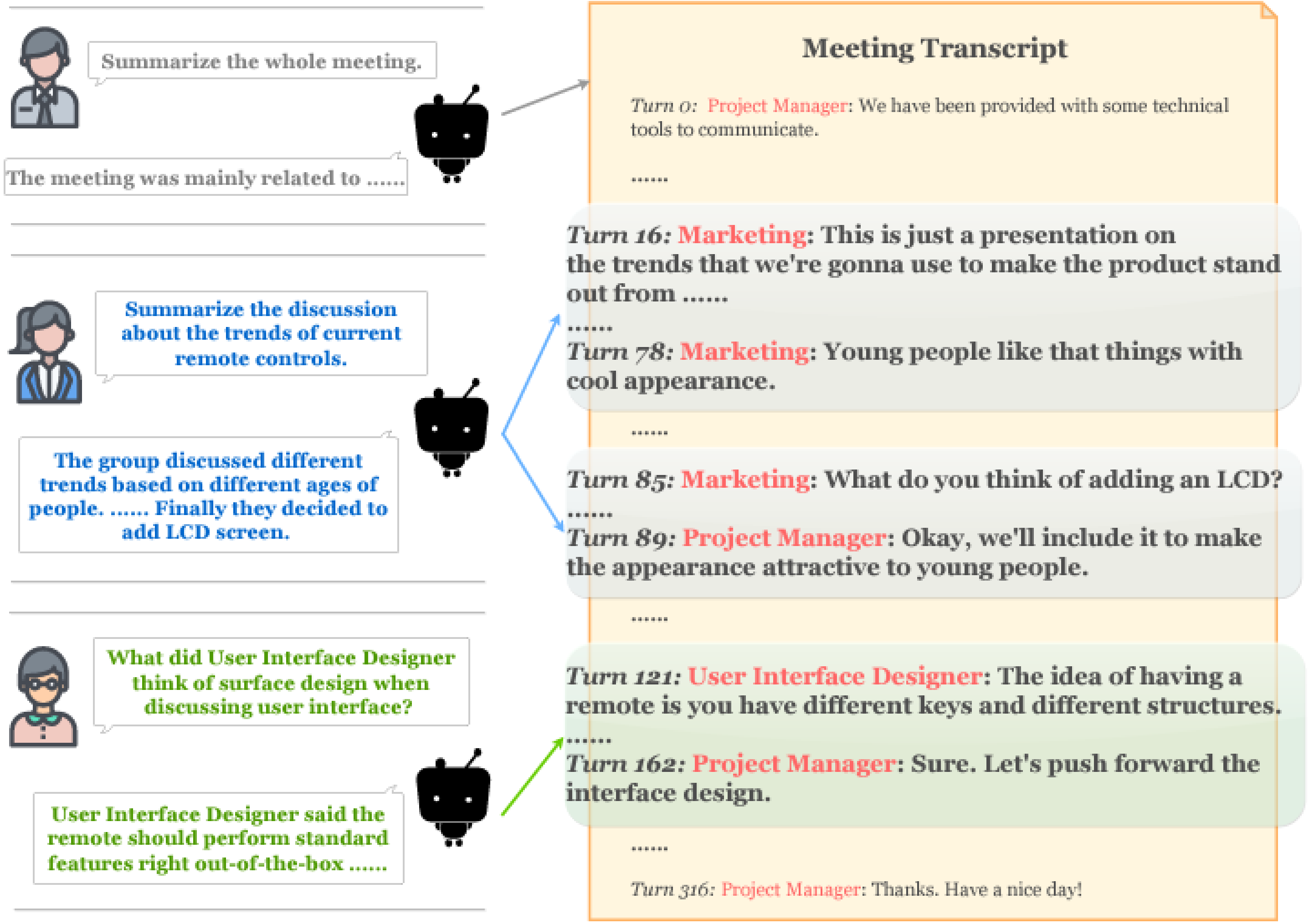
We aim to improve the performance of query-based meeting summarization by introducing a novel approach that leverages the **relationships** between queries.

Our approach involves identifying a relevant query Q_r and its corresponding summary S_r from the history of previously summarized queries for the same meeting. We then feed the model with input in the following format:

$$< s > Q_r < /s > S_r < /s > Q < /s > T < /s >$$

Here, we place the current query in the middle, which is closer to both the main text and the relevant query and corresponding summary. By incorporating relevant queries and summaries from the same meeting, we aim to enhance the quality of the generated summaries and ultimately improve the overall performance of the model. This approach can potentially boost the performance of query-based meeting summarization and lead to more accurate and informative summaries.

QMSum Dataset



- The QMSum benchmark is a recently introduced assessment tool for the task of query-based multi-domain meeting summarization. It comprises a total of 1,808 query-summary pairs extracted from 232 meetings across various domains and has been annotated by human evaluators.
- A query-based meeting summarization application can make work more efficient and help staff understand meetings better in companies.

Zhong et al., NAACL'21

Evaluation

Standard evaluation metrics such as ROUGE

ROUGE measures the precision and recall of the generated summary with respect to the reference summary, using n-gram co-occurrence statistics. Specifically, it computes precision and recall for overlapping n-grams of different lengths, ranging from unigrams (single words) to longer sequences of words.

ROUGU video introduction from HuggingFace

Evaluation Example of ROUGE

Generated summary bigrams

- I really
- really loved
- loved reading
- reading the
- the Hunger
- Hunger Games

Reference summary bigrams

- I loved
- loved reading
- reading the
- the Hunger
- Hunger Games

$$\text{ROUGE-2 recall} = \frac{\text{Num bigram matches}}{\text{Num bigrams in reference}} = \frac{4}{5}$$

$$\text{ROUGE-2 precision} = \frac{\text{Num bigram matches}}{\text{Num bigram in summary}} = \frac{4}{6}$$

$$ROUGE-2\ F1-score = \frac{2(Precision \times Recall)}{(Precision + Recall)}$$

ROUGU video introduction from HuggingFace

Conclusion

To build a query-based meeting summarization system using BART, it will involve

- **Pre-processing:** cleaning and tokenizing the text, and converting it to a format that can be used to fine-tune the BART model. Split the dataset into training, validation, and test sets. User query histories will be part of the input for their future queries.

- **Fine-tuning:** we will experiment with different hyperparameters, such as the learning rate and weight decay to achieve the best possible performance on the query-based meeting summarization task.

- **Evaluating:** we will evaluate its performance on a test set of meetings and their corresponding summaries, using metrics such as ROUGE.

Question?

Thank You!