Yale

Entity-Focused Abstractive Dialogue Summarization

Wai Pan Wong

Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

The growing popularity of text messengers as a means of communication suggests a growing demand in reliable automated summarization. Although much work has been done in the field of linear text summarization, dialogue summarization is considered more difficult – with non-linear conversation flows (questions, discussions), involvement of all personal pronouns which makes entity modeling difficult (first, second, third), and the informality of dialogues (slangs and abbreviations). Compared to extractive summarization, abstractive summarization is more suitable for dialogue summarization. The current state-of-the-art (SOTA) model is BART (which has generation capabilities), a denoising autoencoder that generalizes from the canonical Bidirectional Encoder Representations from Transformers (BERT). BART first corrupts inputs with arbitrary noise function and then learns a reconstruction of the original text. This project is to evaluate the effectiveness of BART on abstractive dialogue summarization, especially on entity modeling, as well as other aspects.

Materials and Methods

The Samsung Abstractive Messenger Summarization (SAMSum) Corpus dataset is used in this project. It consists of 16k dialogues resembling typical messenger conversations scenarios and corresponding abstractive summaries annotated by linguists fluent in English. The dataset is first preprocessed by running byte-pair processing using GPT-2 byte-pair encoding and running binarization for efficient reading and writing of data. Next, the BART large pre-trained model is fine-tuned, with 20,000 total updates and 500 warmup updates, parameters typical for learning rate scheduling. After fine-tuned BART converges, inference is runned on SAMSum corpus by beam search decoding with beam size of 4, with modified maximum and minimum decoding length to account for varying summary sizes. To evaluate the effectiveness of the BART (the pretraining model) model on dialogues, various Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics are calculated. Then manual content-based and linguistic-based evaluations are conducted to check the accuracy and the readability of summaries to human readers.

Table 1. Comparing ROUGE metrics of abstractive summarization. Between BART and Convolutional Seq2Seq model

	BART on SAMSum	SAMSum previous SOTA (convolutional seq2seq model)	Percentage improvement in ROUGE score over previous SOTA
ROUGE-1	50.01	45.41	10.1%
ROUGE-2	25.16	20.65	21.8%
ROUGE-L	51.46	41.45	24.1%

Table 2. An example showing a content error: entity swap in automated summary generated by BART.

DIALOGUE	EXPERT	AUTOMATED
Mary: hey, im kinda broke, lend me a few box Carter: okay, give me an hour, im at the train station Mary: cool, thanks	Mary wants Carter to lend her a few **boxes**. Carter will do it in an hour.	Mary ran out of **money**. Carter is going to lend her some in an hour.

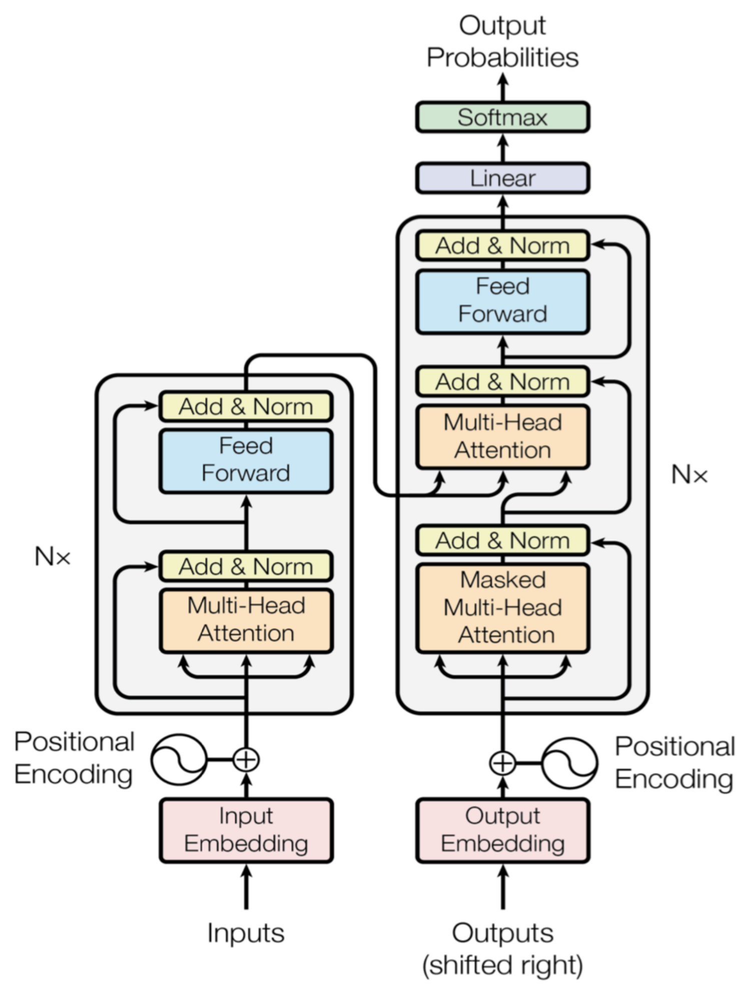
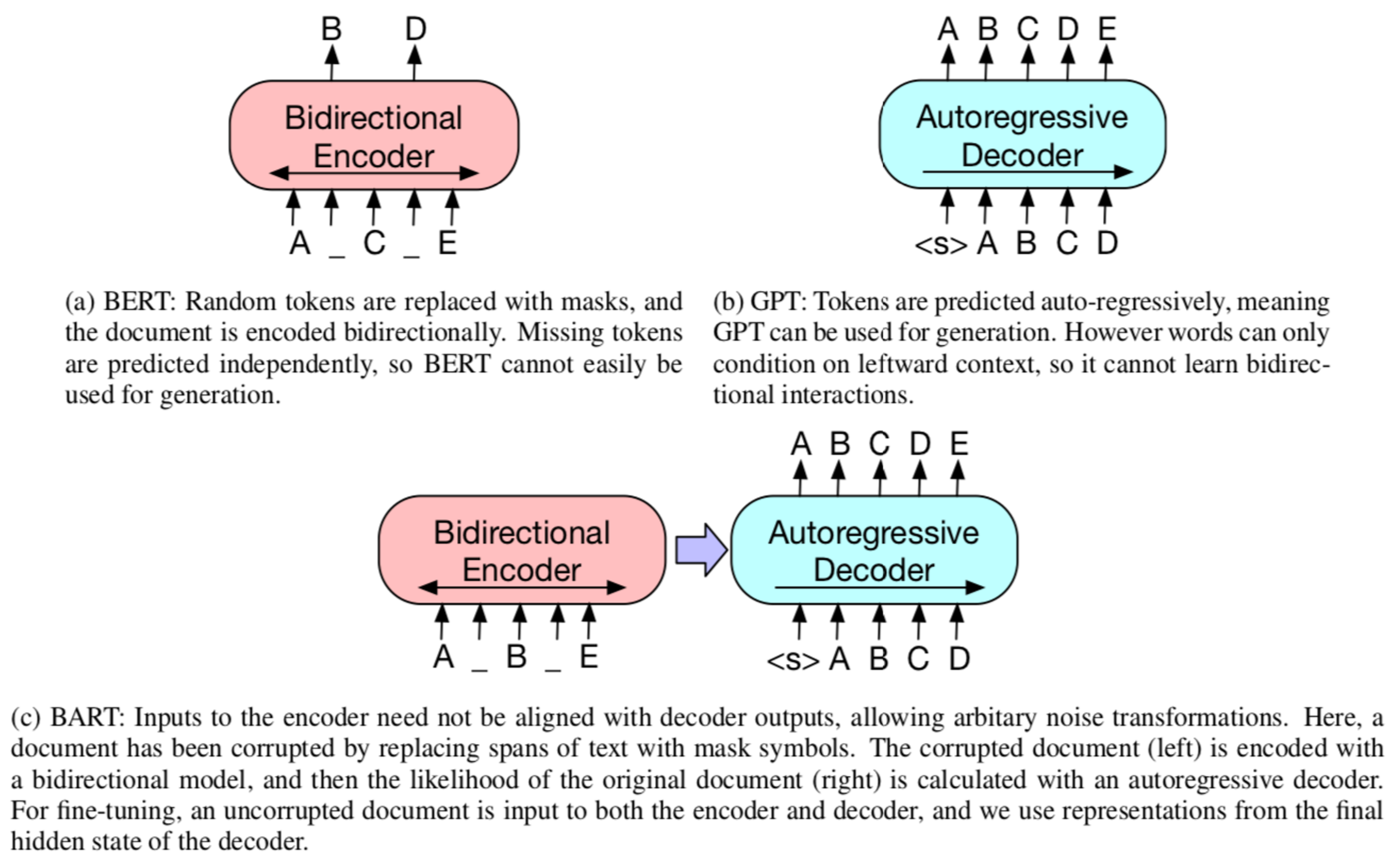


Figure 1: The Transformer - model architecture.

Figure 1. The transformer architecture of Bidirectional Encoder Representations from Transformers (BERT)

https://arxiv.org/pdf/1810.04805.pdf



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure 2. Bidirectional and Auto-Regressive Transformers (BART) as the Generalization of BERT

https://arxiv.org/pdf/1910.13461.pdf

Results

Running with parameters that account for hardware restrictions, the BART pre-training model converges after 10 epochs of training. ROUGE-1 improves by 10.1%, while ROUGE-2 improves by 21.8% and ROUGE-L improves by 24.1%. BART is performing significantly better than the previous SOTA model (convolutional seq2seq model) previously suggested in the SAMSum paper. Rouge-L, which encodes the sentence structure especially the order of word tokens, receives the biggest improvement. So BART is exceptionally good at capturing sentence level structure. However, ROUGE score has its own limitations as it does not take into account human readability so a human evaluation is necessary as a supplement analysis. First, summaries are evaluated in terms of content accuracy, and pronoun swaps and entity swaps are present. A closer analysis suggests that entity swap may result from the frequency of certain words in pretraining, while pronoun swap suggests that the model is not able to separate the topic between who is actually involved in dialogue. In addition, in terms of linguistic quality BART is not totally satisfactory as there are deficiencies in grammaticality, non-redundancy, referential clarity, focus and coherence in the dialogue.

Conclusion & Further Investigations

The deficiencies of BART demonstrated in analysis suggests further modifications are necessary for BART to better match human readability. First, an improved metric could take into account more than word overlap without having to rely on human judgement. For example, the metric CIC (critical information loss) focuses on important entities in a dialogue. On the other hand, some entity module should take care of entities and understand the relationship between them. One of the papers suggests SENECA, which makes use of entity information to select relevant information. It would be promising to evaluate BART with the above metric as well as include an entity module to see the further improvements.

Acknowledgement

I would like to thank Professor Dragomir Radev and Alexander Fabbri for their guidance throughout the project.