Yale

An Analysis of Long Dialogue Summarization

Ansong Ni*, Yusen Zhang*, Tao Yu, Rui Zhang, Dragomir Radev

LILY Lab

Introduction

Dialogues, such as meetings, interviews, and podcasts are typically long and verbose, thus a summarization model can help the readers capture important information and improve their reading efficiency. However, the length of such dialogues typically exceeds the inputs limits imposed by recent transformer-based models, making it difficult to train an end-to-end summarization model for longer dialogues. Moreover, the interactive nature of dialogues makes it more context-dependent and information more sparsely distribution than articles, thus more challenging to generate a succinct summary. In this work, we perform comprehensive analysis on long dialogue summarization problem. To address the lengthy input problem, we investigate extended transformer models as Longformer, and several dialogue utterance retrieval methods for a \textit{retrieve-then-summarize} pipeline model, as well as hierarchical dialogue encoding models. We find that pipeline models that use a dialogue utterance retrieval model yields the best performance, and the summary quality can be further improved with a stronger retrieval model. We further find that pretraining on external summarization datasets can effectively improve the performance of dialogue summarization models.

Datasets

QMSum: A query-based meeting dialogue summarization dataset. Dataset size is 1.6K and the dialogues are around 9K tokens long with an average 9 speakers on each dialogue. Gold spans are annotated for the dialogue utterance retrieval/locating task.

SummScreen: A dialogue summarization dataset from TV show transcripts and fan-generated recaps from two different websites. Around 6K tokens for each dialogue with 27K data examples. The summaries are longer than QMSum, reaching an average of 330 tokens.

Retrieval-Summarize Framework

Baseline Retrieval Models:

	ROUGE-1	ROUGE-2	ROUGE-L
Random	31.1	7.9	20.9
Cutoff	32.6	8.7	21.6
TF-IDF	32.5	8.5	21.4
BM25	32.9	9.0	22.0
Gold Span	**38.9**	**13.9**	**26.3**

Table 1. Summarization Performance on QMSum after Retrieval

Observation:
1) Important to keep continuity of dialogues;
2) Good retrieval model could help summarization.

BERT-based Segment Relevance Modeling:

Methodology:
1) Divide each dialogue into segments using sliding window (size 10, stride 5)
2) Input: [CLS] query [SEP] concat-utts [SEP]
3) Model: BERT/RoBERTa + softmax => binary label

	bert-base	Roberta-base + weight scaling
Acc	87.5	84.0
Pos-Recall	35.6	56.0
Pos-Precision	38.6	33.6
Pos-F1	37.0	42.0
Pos-AUC	77.9	81.6

Table 2. BERT-based Retrieval Results

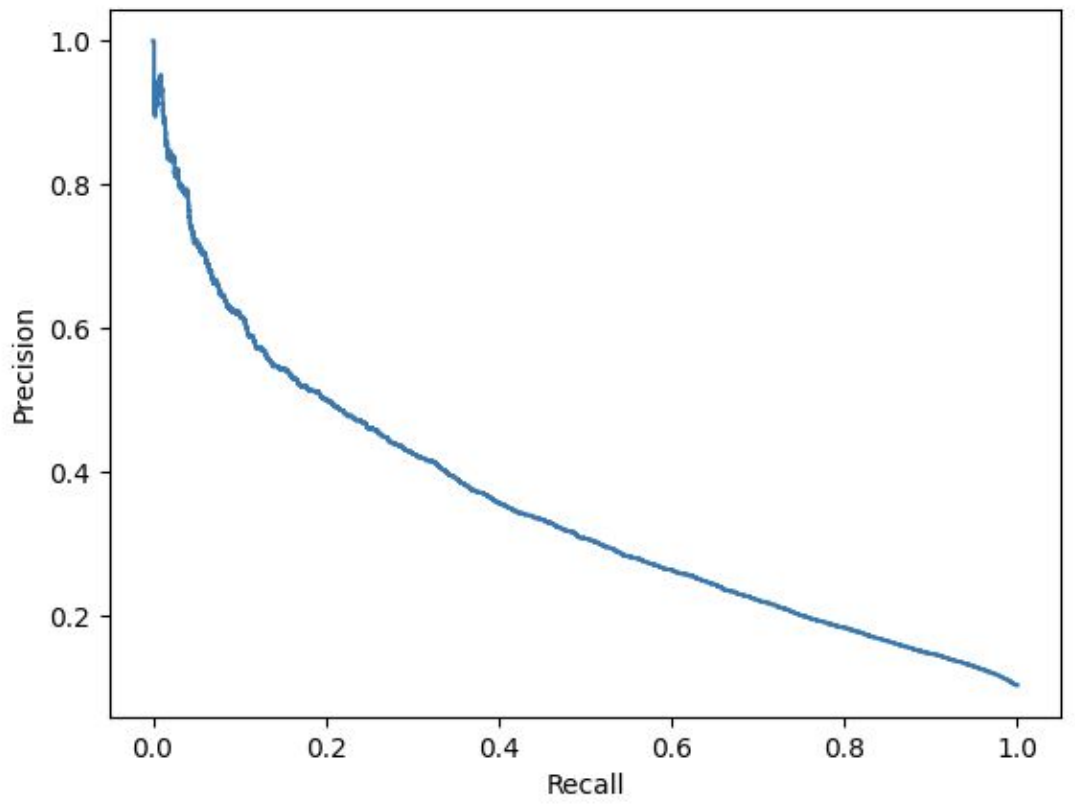

Figure 1. Precision-Recall Curve

No Retrieval (End-to-end) Model

Methods:
1) **BART -** with cutoff at max 1024 tokens;
2) **BART-extended -** extended positional embedding to take 2048 tokens;
3) **Longformer -** Use sparse (along diagonal) attention matrix + global attention to perform more efficient learning, able to take as long as 16K tokens length input

	max_input_len	ROUGE-1	ROUGE-2	ROUGE-L
bart-large-cnn	1024	38.7	14.3	26.4
bart-large-cnn (extended)	2048	37.8	13.1	25.2
Longformer-large-16384	8192	35.9	11.9	24.2