Dialogue Summarization: Datasets and Pretraining Strategies

Yang Liu
Microsoft

Motivation

Assume we finished the meeting……

Cool! I can enjoy my weekend!

Emm, what did we discuss in the meeting…?

Motivation

Assume we finished the meeting……

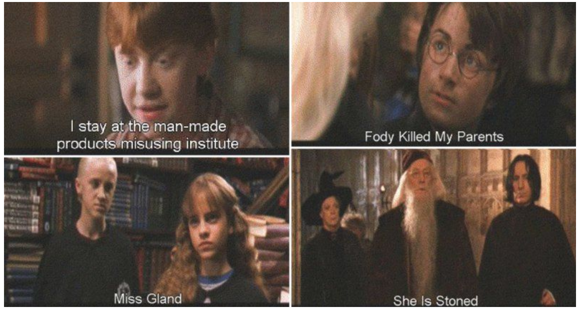
Cool! I can enjoy my weekend!

Emm, what did we discuss in the meeting…?

Dialogue summarization!!

Motivation

An increasing number of dialogues are recorded and transcribed



Microsoft Teams

zoom

Motivation

- Existing research on text summarization focuses on **monologic** texts.

- **Dialogue**, as an important communicative channel, has received significantly less attention.

Roadmap

1. Datasets for Dialogue Summarization

2. DialogLM, a strong pretrained model

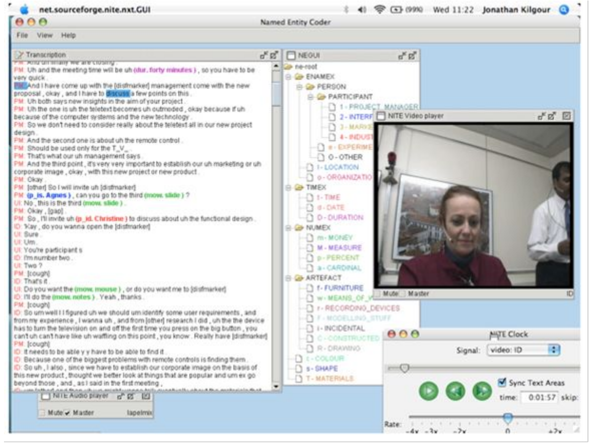
3. Challenges and Future works

The Challenge of Building a Dialogue Summarization Dataset

- Dialogues that we want to summarize are hard to get, and hard to publicize
- An informative conversation could last for 1-2 hours, leading to over 20k tokens, annotation could be costly

Existing Dialogue Summarization Datasets

- Current research on dialogue summarization mainly uses AMI/ICSI and SAMSum datasets.



(a) AMI meeting dataset

1. ashleigh: looks like we're going to the cinema!!
2. ashleigh: <file_gif>
3. peter: you got the job??
4. ashleigh: i got hte job! :d
5. peter: <file_gif>
6. ashleigh: <file_gif>

(b) SAMSum dataset

Existing Dialogue Summarization Datasets

- AMI meeting corpus
  - Consisting of 137 virtual multi-party meetings
  - Limited to its scale $\nRightarrow$ neural models

- SAMSum
  - A written online conversation summarization dataset
  - Conversations are too short
  - Language style differs from real-scenario dialogues, mainly about leisure chitchats

Dialogue Summarization Datasets in the New Era

- DialogSum
  - Real-Life Scenario Dialogue Summarization
- MediaSum
  - Media Interview Dataset for Dialogue Summarization
- QMSum
  - Query-based Multi-domain Meeting Summarization

DialogSum

Real-Life Scenario Dialogue Summarization

A Comparison between the Real-Life Scenario Dialogue and Online Chit-Chat

(a) Dialogue from DIALOGSUM:
#Person_1#: Good morning. I wonder whether you have got an answer from your superior.
#Person_2#: Yes, we had a meting about it yesterday afternoon.
#Person_1#: What's the answer?
#Person_2#: We decided that we could agree to your price, but we are a bit worried about the slow delivery.
#Person_1#: Let me see. I quoted your delivery in three months, didn't I?
#Person_2#: Yes, but we hope that the wool could reach us as soon as possible.
#Person_1#: I thought you would. So I rang Auckland last night. As you are our biggest customer, they agreed to ship the order on the first vessel available that will leave Auckland next month.
#Person_2#: Good, if you agree we'll draft the agreement right away and sign it then.
#Person_1#: By all means.

Summary from DIALOGSUM: #Person_1# and #Person_2# agree to sign an agreement ***since*** #Person_1# could speed up the delivery as #Person_2# hopes.

(b) Dialogue from SAMSum:
…
Leo: BTW what are those pics?
Ryan: Pics from Italy!!! :):):):)))))))))
Leo: Yeah. They seem nice. (‘A`)
Ryan: That's all???? I need more reactions!!!!!!!!!!
Leo: I'm tied to this office and working like a slave. AM I SUPPOSED TO SAY \"I AM SO JEALOUS!!!!!!!!\"?😬😬😬
…

Summary from SAMSum: Ryan is in Italy while Leo is working hard and wishing he could win the lottery.

DialogSum

Real-Life Scenario Dialogue Summarization

- Dialogues are from 3 public datasets, DailyDialog, DREAM and MuTual, and an online English practice website.
  - Under rich real-life scenarios, including diverse task-oriented scenarios
  - Multi-turn dialogues within reasonable lengths
- Annotated by Human Annotators
  - Compression rate: 15%~20%
  - Written from an observer perspective

Datasets	Lan. style	Domain	Scenario	Dialogs	Data size	#Tokens/dial.	#Tokens/turn	#Comp. rate
AMI	spoken	single	meeting	137	100hrs (video)	4,757	16.5	0.07
SAMSum	written	multiple	online	16,369	1.5M (token)	94	8.4	0.30
DIALOGSUM	spoken	multiple	daily life	13,460	1.8M (token)	131	13.8	0.18

DialogSum
Real-Life Scenario Dialogue Summarization

Inter-annotator agreement is reasonable

Human Annotated Summary	R1	R2	RL
Summary1 to Summary2	52.90	26.01	50.42
Summary1 to Summary3	53.85	27.53	51.65
Summary2 to Summary3	53.30	26.61	50.44
Average	53.35	26.72	50.84

A bit more challenging than SAMSum

Model	CNNDM			XSum			SAMSum			DIALOGSUM		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Transformer	40.21	17.76	37.09	29.41	9.77	23.01	37.20	10.86	34.69	35.91	8.74	33.50
$\text{UNILMv2}_{\text{BASE}}$	43.16	20.42	40.14	44.00	21.11	36.08	50.53	26.62	48.81	47.04	21.13	45.04
$\text{BART}_{\text{LARGE}}$	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**	**53.12**	**27.95**	**49.15**	47.28	21.18	44.83

DialogSum is too short, I want **more challenging data**!

MediaSum

Media Interview Dataset for Dialogue Summarization

- Interview transcripts from NPR/CNN
  - NPR has a editor-written summaries
  - CNN has a list of discussed topics

Summary: The ’sea rocket’ shows preferential treatment to plants that are its kin. Evolutionary plant ecologist Susan Dudley of McMaster University in Ontario discusses her discovery.

A: This is Day to Day. I’m Madeleine Brand.
B: And I’m Alex Cohen.
A: Coming up, the question of who wrote a famous religious poem turns into a very unchristian battle.
B: First, remember the 1970s? People talked to their houseplants, played them classical music. They were convinced plants were sensuous beings and there was that 1979 movie.
… …
A: OK. Thank you.
B: That’s Susan Dudley. She’s an associate professor of biology at McMaster University in Hamilt on Ontario. She discovered that there is a social life of plants.

MediaSum

Media Interview Dataset for Dialogue Summarization

- Interview transcripts from NPR/CNN
  - NPR has a editor-written summaries
  - CNN has a list of discussed topics

Statistics	NPR	CNN
Dialogues	49,420	414,176
Avg. words in dialogue	906.3	1,630.9
Avg. words in summary	40.2	11.3
Turns	24.2	30.7
Speakers	4.0	6.8
Novel summary words	33.6%	24.9%

MediaSum
Media Interview Dataset for Dialogue Summarization

- Interview transcripts from NPR/CNN
  - NPR has a editor-written summaries
  - CNN has a list of discussed topics

Model	R-1	R-2	R-L
LEAD-3	14.96	5.10	13.29
PTGen	28.77	12.24	24.18
UniLM	32.70	17.27	29.82
BART	**35.09**	**18.05**	**31.44**

MediaSum is interviews, I want **more practical dialogues**!

It makes no sense to summary a long dialogue with **just one summary**!

QMSum
Query-based Multi-domain Meeting Summarization

People are interested in various topics in a single meeting.

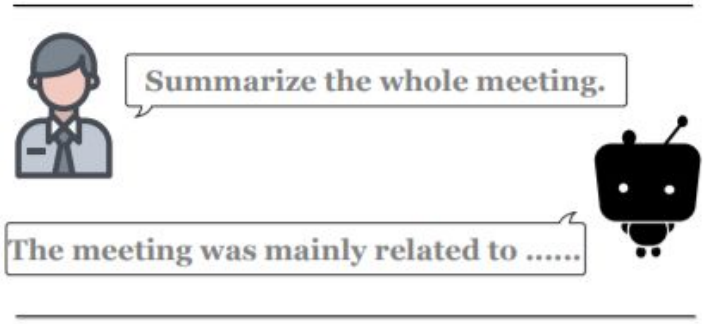
What did the group members say when discussing X?
What was A's opinion towards X?
What was the conclusion about X?

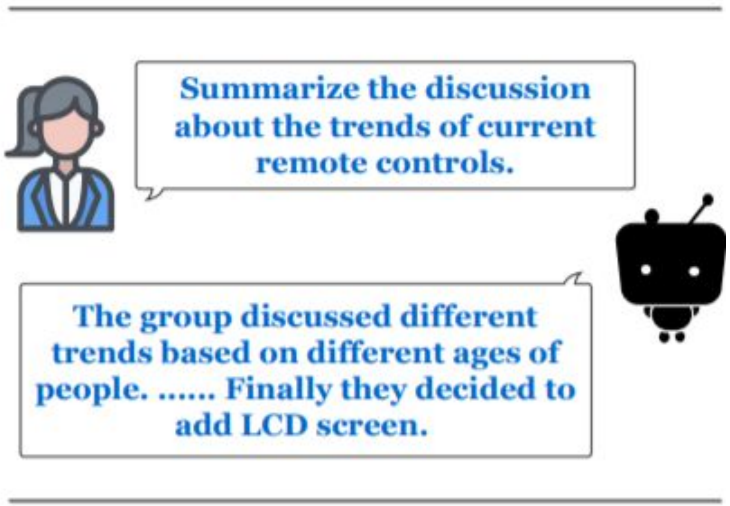
Why not make a meeting summarization benchmark with these questions?

QMSum

Query-based Multi-domain Meeting Summarization

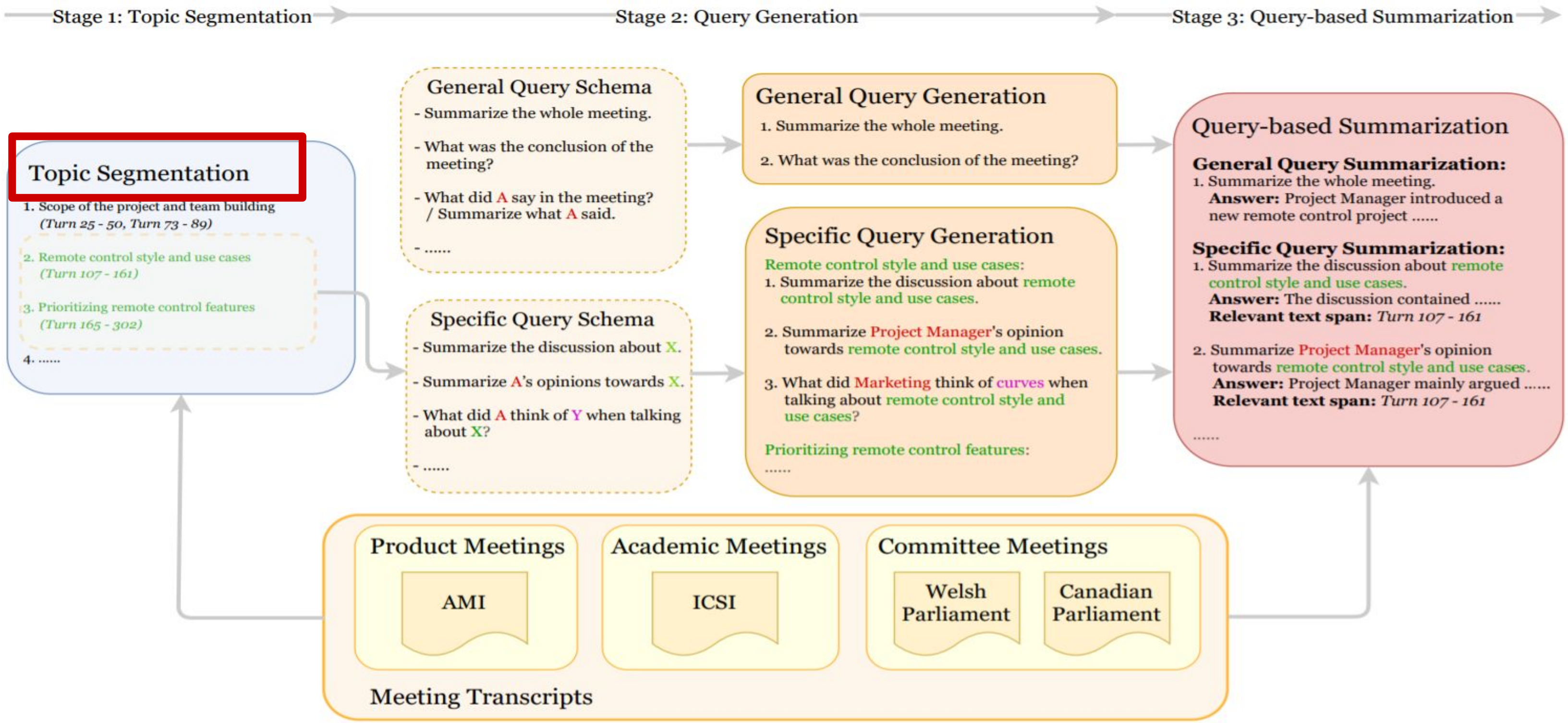
- **Query-based Meeting Summarization (QMSum)**
  - Given a meeting script and a query, summarize the relevant contents to answer the query.

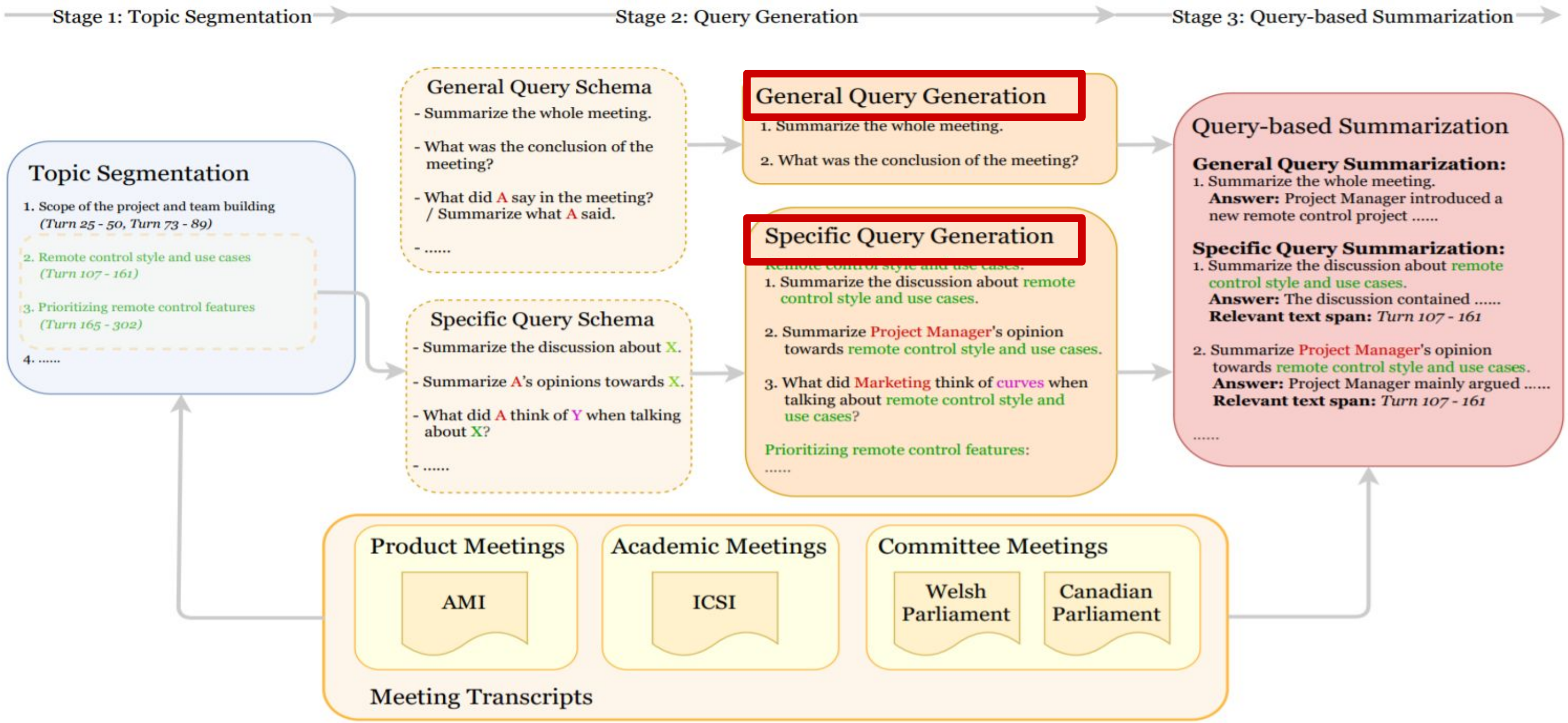




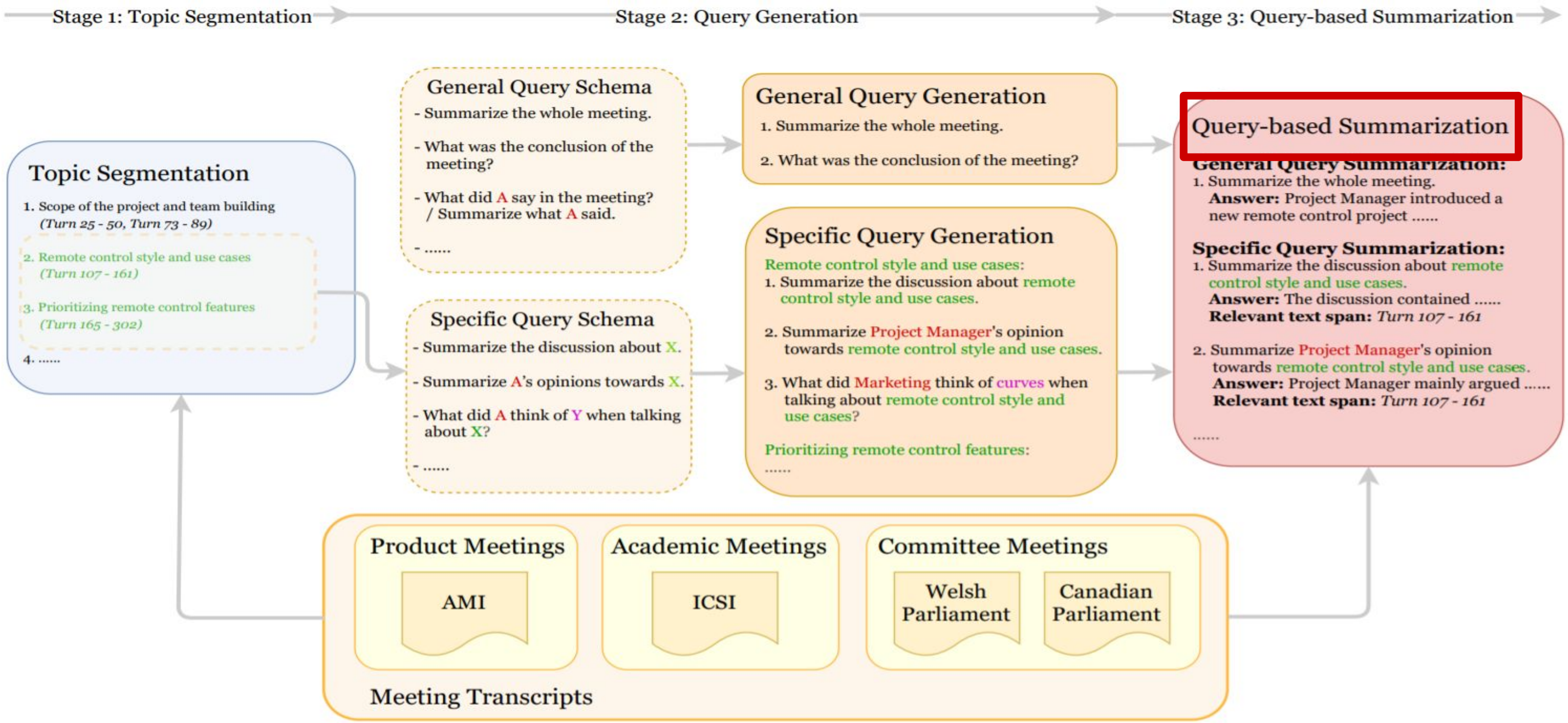
QMSum
Query-based Multi-domain Meeting Summarization

- Multi-domain Meeting Collection
  - Product Meeting
    - AMI (Carletta et al., 2005)
  - Academic Meeting
    - ICSI (Janin et al., 2003)
  - Committee Meeting
    - Welsh Parliament
    - Canadian Parliament


Stage 1: Topic Segmentation
Stage 2: Query Generation
Stage 3: Query-based Summarization
Topic Segmentation
1. Scope of the project and team building
(Turn 25 - 50, Turn 73 - 89)
2. Remote control style and use cases
(Turn 107 - 161)
3. Prioritizing remote control features
(Turn 165 - 302)
4. ......
General Query Schema
- Summarize the whole meeting.
- What was the conclusion of the meeting?
- What did A say in the meeting? / Summarize what A said.
- ......
Specific Query Schema
- Summarize the discussion about X.
- Summarize A's opinions towards X.
- What did A think of Y when talking about X?
- ......
General Query Generation
1. Summarize the whole meeting.
2. What was the conclusion of the meeting?
Specific Query Generation
Remote control style and use cases:
1. Summarize the discussion about remote control style and use cases.
2. Summarize Project Manager's opinion towards remote control style and use cases.
3. What did Marketing think of curves when talking about remote control style and use cases?
Prioritizing remote control features:
......
Query-based Summarization
General Query Summarization:
1. Summarize the whole meeting.
Answer: Project Manager introduced a new remote control project ......
Specific Query Summarization:
1. Summarize the discussion about remote control style and use cases.
Answer: The discussion contained ......
Relevant text span: Turn 107 - 161
2. Summarize Project Manager's opinion towards remote control style and use cases.
Answer: Project Manager mainly argued ......
Relevant text span: Turn 107 - 161
......
Product Meetings
AMI
Academic Meetings
ICSI
Committee Meetings
Welsh Parliament
Canadian Parliament
Meeting Transcripts



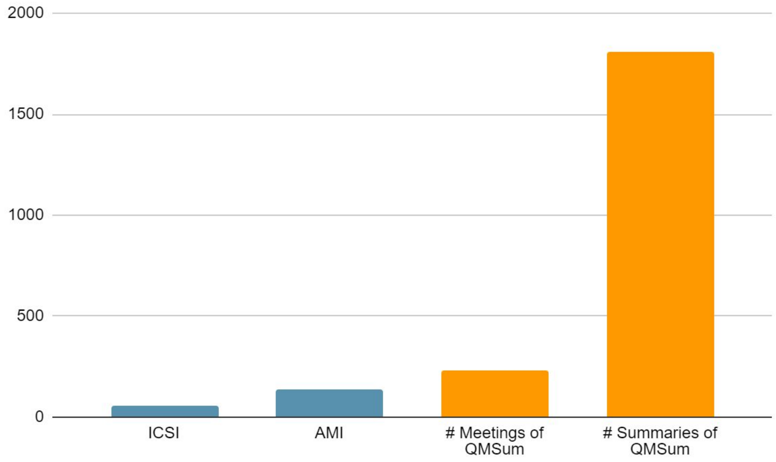
Stage 1: Topic Segmentation
Stage 2: Query Generation
Stage 3: Query-based Summarization
Topic Segmentation
1. Scope of the project and team building
(Turn 25 - 50, Turn 73 - 89)
2. Remote control style and use cases
(Turn 107 - 161)
3. Prioritizing remote control features
(Turn 165 - 302)
4. ......
General Query Schema
- Summarize the whole meeting.
- What was the conclusion of the meeting?
- What did A say in the meeting? / Summarize what A said.
- ......
Specific Query Schema
- Summarize the discussion about X.
- Summarize A's opinions towards X.
- What did A think of Y when talking about X?
- ......
General Query Generation
1. Summarize the whole meeting.
2. What was the conclusion of the meeting?
Specific Query Generation
1. Summarize the discussion about remote control style and use cases.
2. Summarize Project Manager's opinion towards remote control style and use cases.
3. What did Marketing think of curves when talking about remote control style and use cases?
Prioritizing remote control features:
......
Query-based Summarization
General Query Summarization:
1. Summarize the whole meeting.
Answer: Project Manager introduced a new remote control project ......
Specific Query Summarization:
1. Summarize the discussion about remote control style and use cases.
Answer: The discussion contained ......
Relevant text span: Turn 107 - 161
2. Summarize Project Manager's opinion towards remote control style and use cases.
Answer: Project Manager mainly argued ......
Relevant text span: Turn 107 - 161
......
Product Meetings
AMI
Academic Meetings
ICSI
Committee Meetings
Welsh Parliament
Canadian Parliament
Meeting Transcripts



Stage 1: Topic Segmentation
Stage 2: Query Generation
Stage 3: Query-based Summarization
Topic Segmentation
1. Scope of the project and team building (Turn 25 - 50, Turn 73 - 89)
2. Remote control style and use cases (Turn 107 - 161)
3. Prioritizing remote control features (Turn 165 - 302)
4. ......
General Query Schema
- Summarize the whole meeting.
- What was the conclusion of the meeting?
- What did A say in the meeting? / Summarize what A said.
- ......
Specific Query Schema
- Summarize the discussion about X.
- Summarize A's opinions towards X.
- What did A think of Y when talking about X?
- ......
General Query Generation
1. Summarize the whole meeting.
2. What was the conclusion of the meeting?
Specific Query Generation
Remote control style and use cases:
1. Summarize the discussion about remote control style and use cases.
2. Summarize Project Manager's opinion towards remote control style and use cases.
3. What did Marketing think of curves when talking about remote control style and use cases?
Prioritizing remote control features:
......
Query-based Summarization
General Query Summarization:
1. Summarize the whole meeting.
Answer: Project Manager introduced a new remote control project ......
Specific Query Summarization:
1. Summarize the discussion about remote control style and use cases.
Answer: The discussion contained ......
Relevant text span: Turn 107 - 161
2. Summarize Project Manager's opinion towards remote control style and use cases.
Answer: Project Manager mainly argued ......
Relevant text span: Turn 107 - 161
......
Product Meetings
AMI
Academic Meetings
ICSI
Committee Meetings
Welsh Parliament
Canadian Parliament
Meeting Transcripts

QMSum
Query-based Multi-domain Meeting Summarization

- QMSum is the currently largest meeting summarization dataset.
- The average length is 9069.8 words.
- Meetings, queries, summaries, main topics, relevant text spans.



With all these data, can we have **one model to summ them all**?

DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization

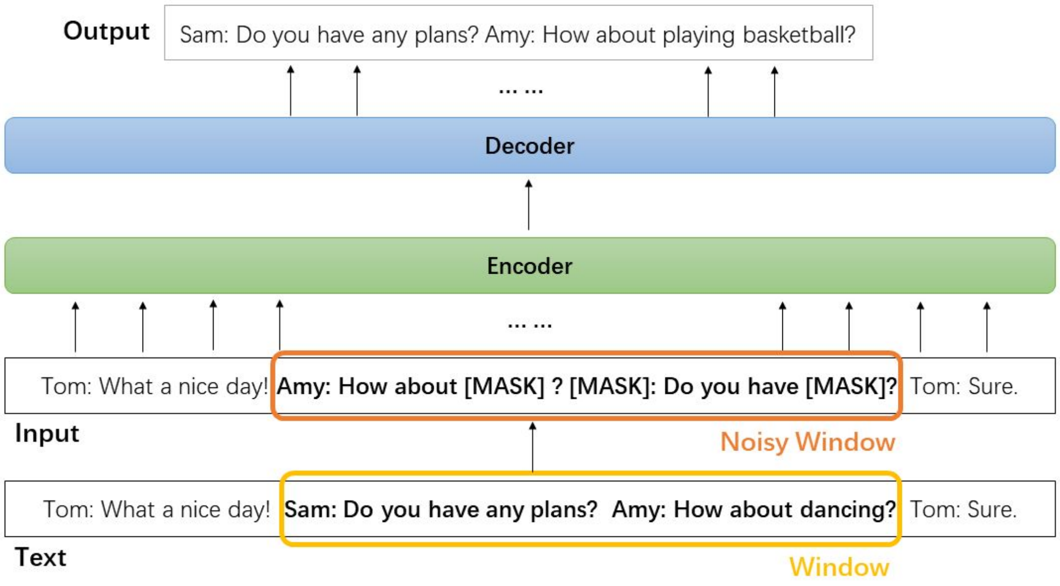
- Lack of powerful tools to process long conversation
  - Dialogue-related pre-trained models focus on several specific tasks
    - Dialogue Response Generation
    - Addressee Selection
    - Response Secltion
  - They can only process short conversations (~100 words, ~10 turns), but can't model long dialogues (> 5,000 words, >300 turns)

DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization

- Lack of powerful tools to process long conversation
  - Pre-trained models for long document doesn't consider the characteristics of the dialogue
    - Longfomer
    - Bigbird
    - …
  - They are not familiar with the special format of the dialogue
    - There are multiple speakers in a long conversation
    - The basic unit of dialogue is "**Turn**" instead of "**Sentence**"
  - **We need a pre-trained model to process various types of long dialogues!**

DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization

Window-based Denoise (Pre-train Task)

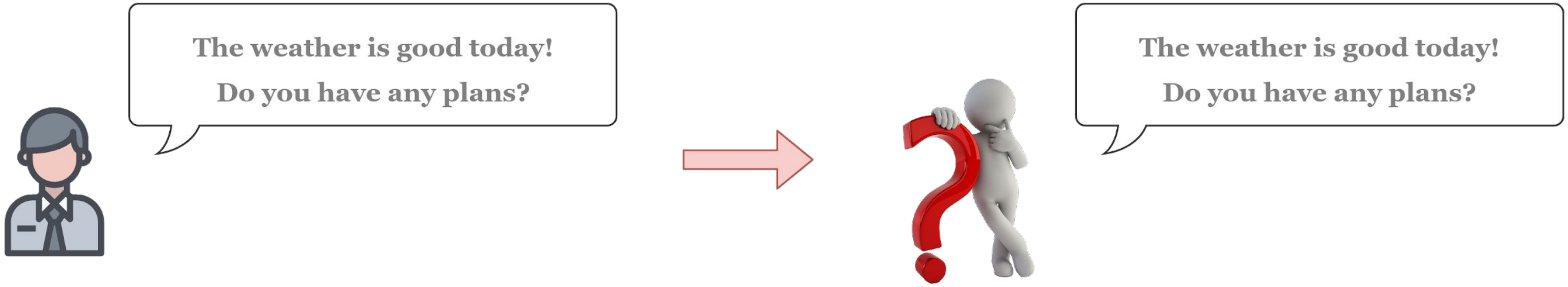


- ❏ Full Text Denoise
  - ❏ BART
  - ❏ Can’t be used for long sequences!
- ❏ Sentence-level Mask
  - ❏ PEGASUS
  - ❏ A single turn may have no useful information
  - ❏ Multiple turns in the dialogue are coherent

Method

★ How to Generate a Noisy Window?

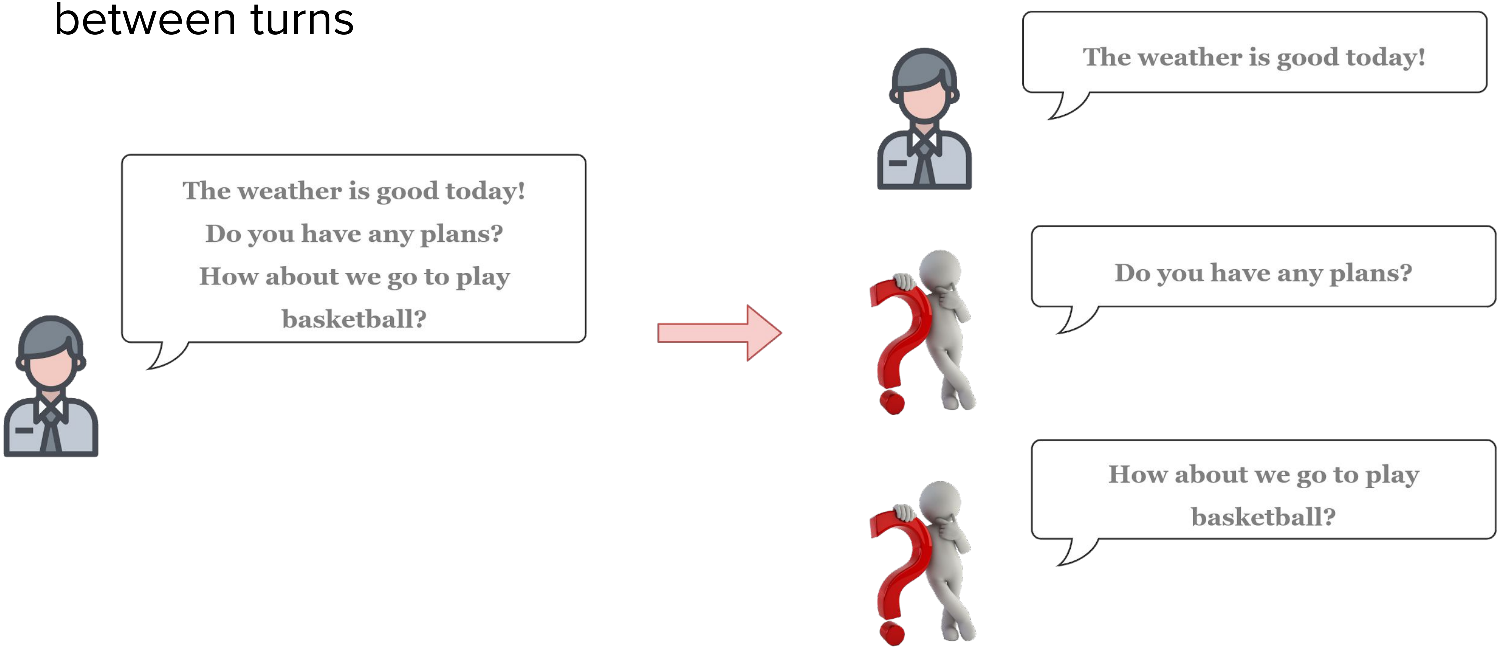
- Noise 1: Speaker Mask
- Denoise it can help the model to identify the speaker



Method

★ How to Generate a Noisy Window?

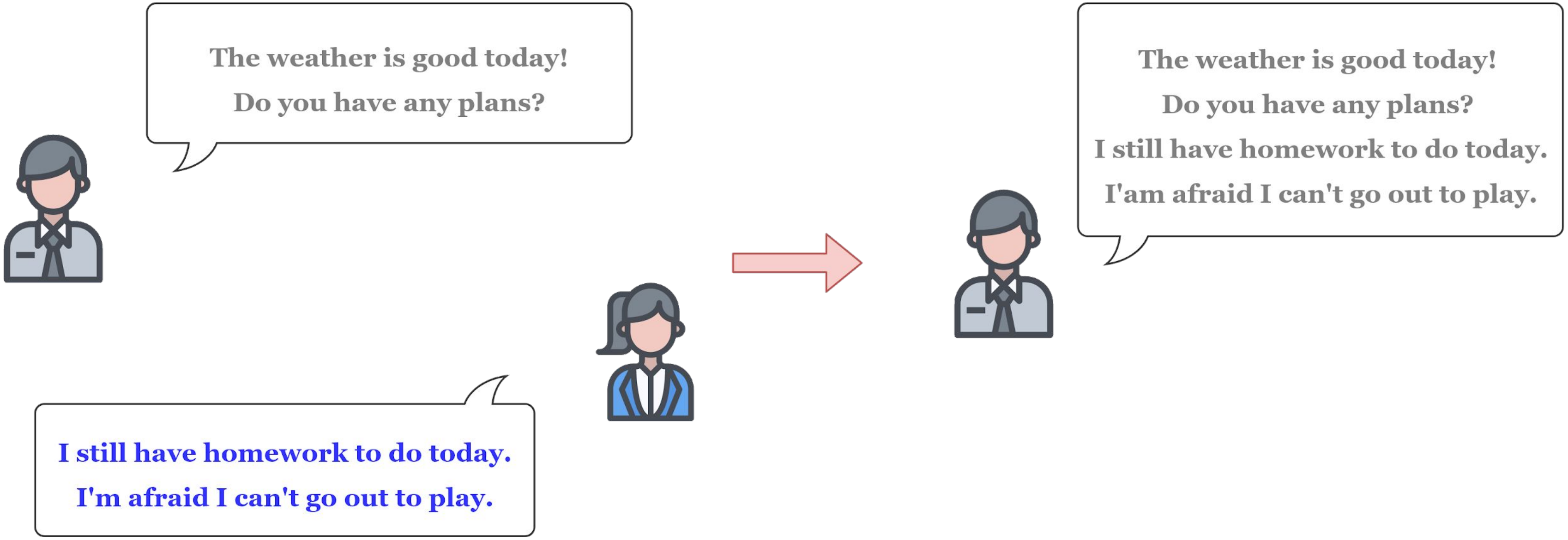
- Noise 2: Turn Splitting
- Denoise it can help the model identify the speaker and the boundary between turns



Method

★ How to Generate a Noisy Window?

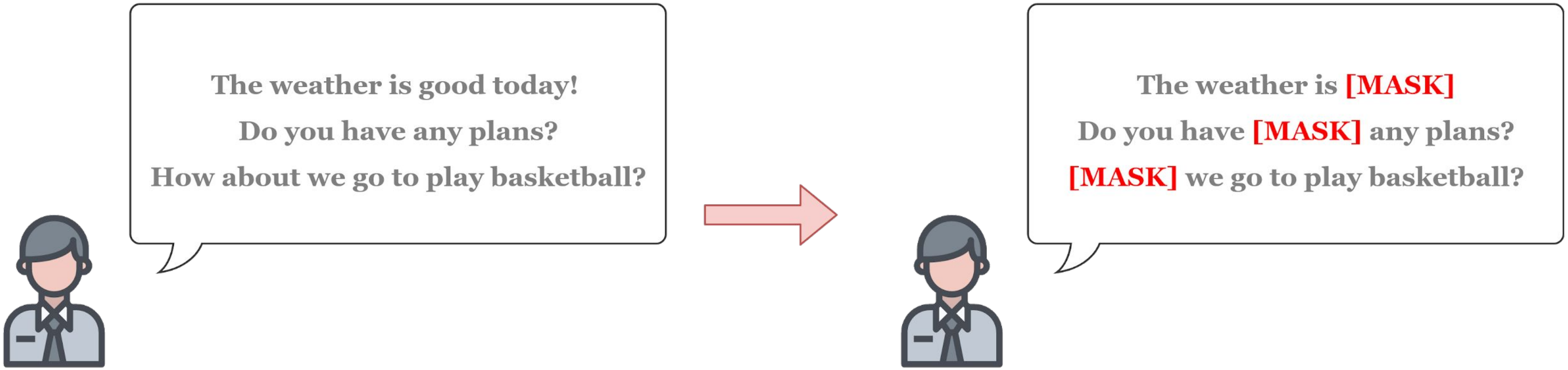
- Noise 3: Turn Merging
- Denoise it can help the model identify the speaker and the boundary between turns



Method

★ How to Generate a Noisy Window?

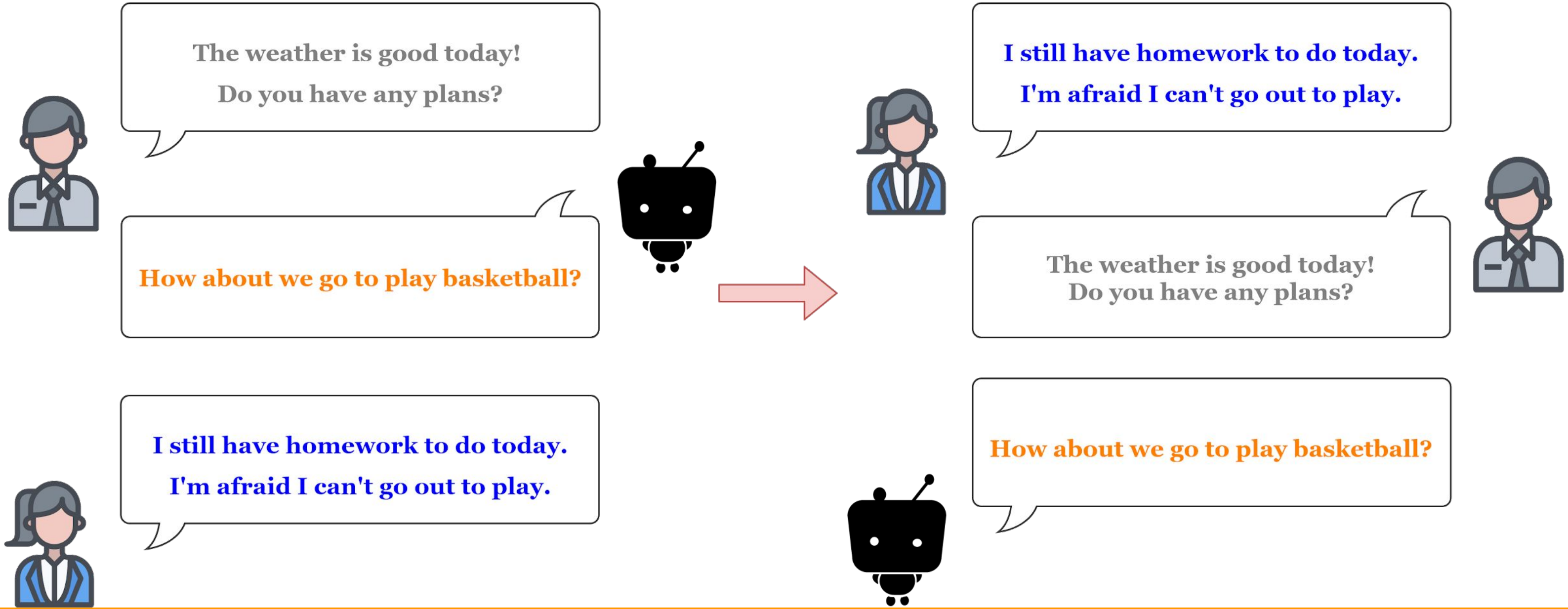
- Noise 4: Text Infilling
- Denoise it can help the model understand the content of the utterance



Method

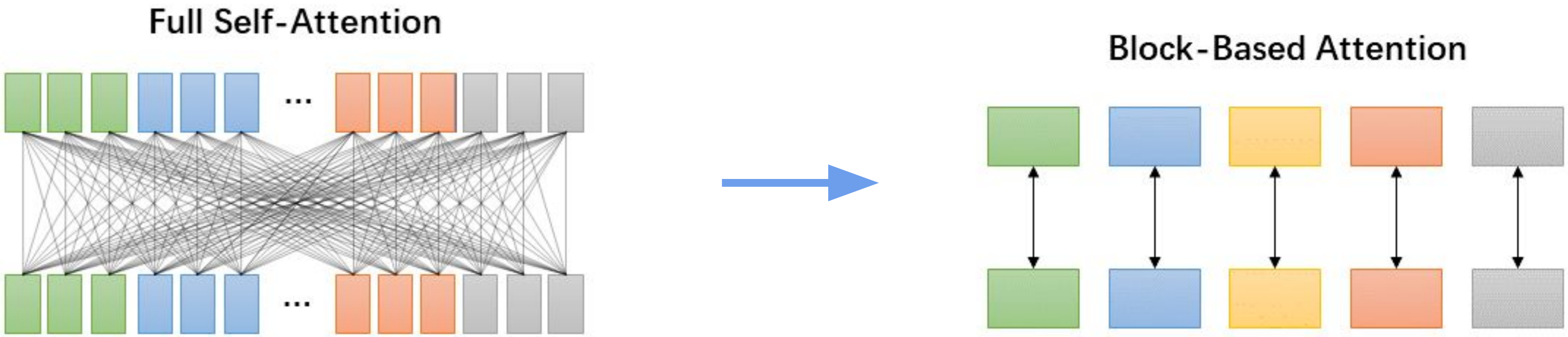
★ How to Generate a Noisy Window?

- Noise 5: Turn Permutation
- Denoise it can help the model understand the order of turns in the dialogue

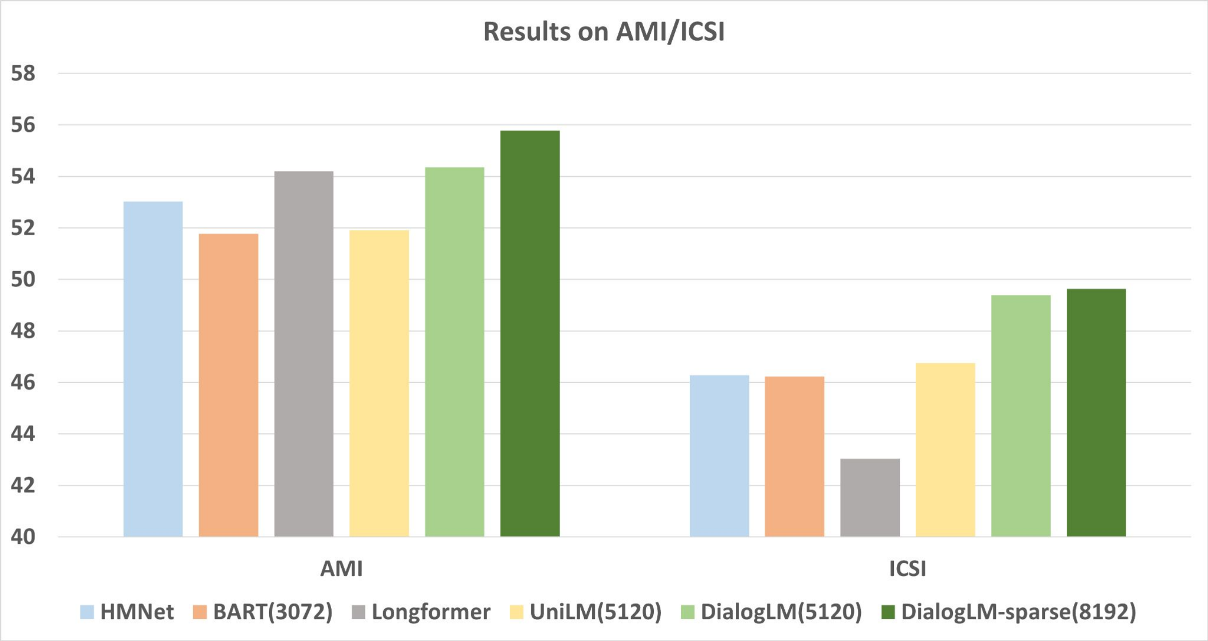


Method

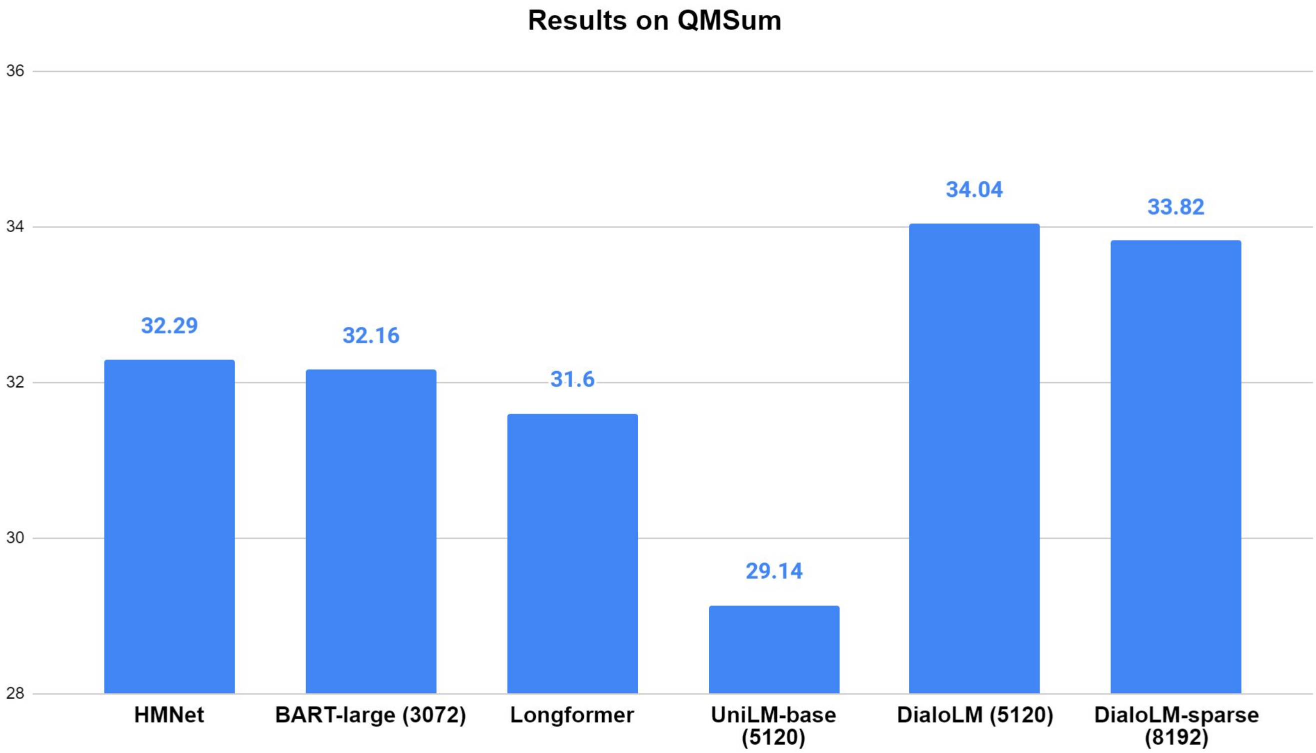
- ★ Model Architecture for **DialoLM**
  - Backbone Model: UniLM
  - Limitations
    - Only 512 words can be processed
    - No pre-training for dialogue
  - Introduc sparse attention to input longer text and reduce training time
    - Full Self-Attention → Block-based Attention



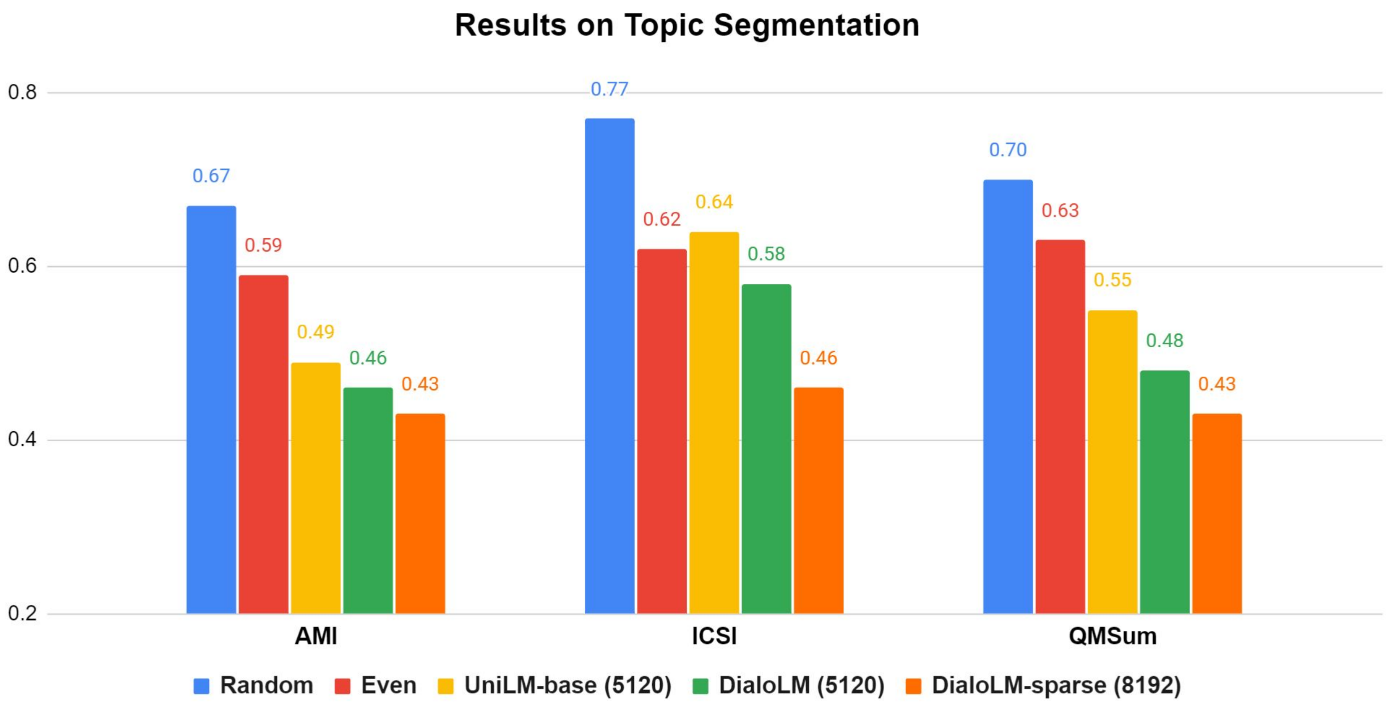
Experiments



Experiments



Experiments



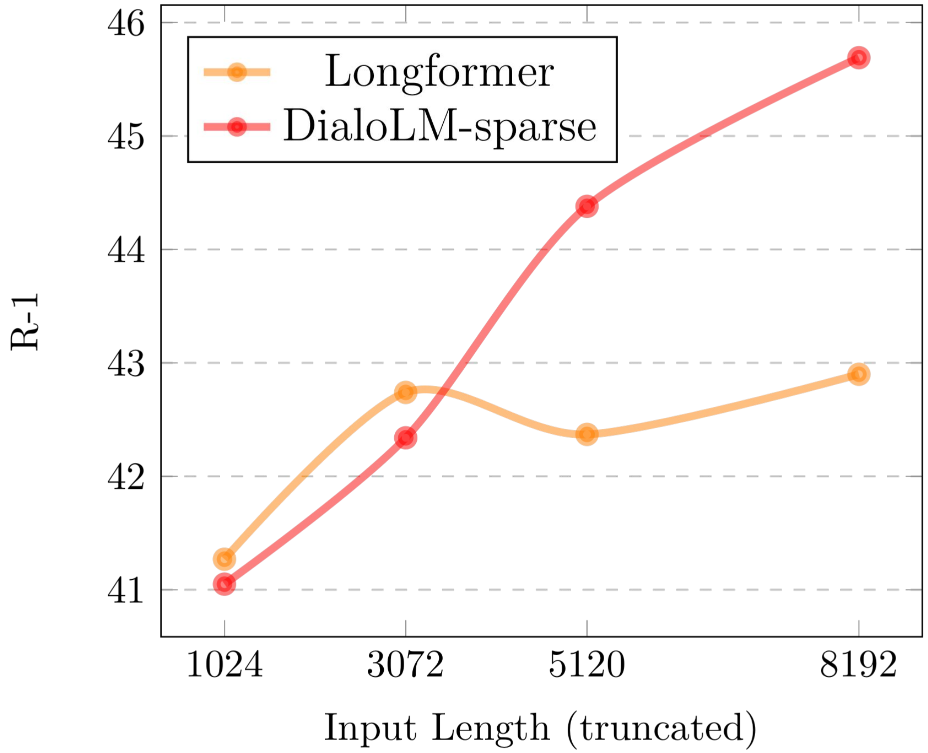
Experiments

Ablation study shows turn splitting/merging is the most important objective

Model	QMSum
DialogLM-sparse	33.69
- Sparse Attention	**34.02**
- Pre-train	29.14
- Speaker Mask	33.52
- Turn Splitting / Merging	32.76
- Text Infilling	33.27
- Turn Permutation	33.22

Analysis

Influence of Input Sequence Length



- ❑ Increasing the input length does not significantly improve Longformer
- ❑ DialoLM is capable of processing long dialogues. The longer the input, the better the performance

Challenges

- Faithfulness and Hallucination are still the biggest problem of abstractive summarization
  - For DialogLM on QMSum
    - 74% generated summaries contain factual errors!
    - 31% generated summaries are completely unrelated to the given query!

Contextual/Discourse Reasoning

DIALOGUE - D:
#Person_1#: Hey, don't I know you from somewhere?
#Person_2#: No, sorry. I don't think so.
#Person_1#: Didn't you use to work at Common Fitness Gym?
#Person_2#: No, I'm afraid I did not.
#Person_1#: Oh, but I know you from somewhere else. Did you use to work at the movie theater downtown? You did. Yes. It's you. I go there all the time and you always sell me popcorn and soda.
#Person_2#: No, that's not me either. Sorry, ma'am. Perhaps I look familiar to you, but ...
#Person_1#: No, I know you. I have met you before! Hold on. Let me think. This is driving me crazy. I know that we've talked before. Oh, I remember now. You work at the Whole Bean Cafe on the corner. It that right?
#Person_2#: No, wrong again. Sorry, ma'am, but I really have to get going.

SUMMARY – D1: #Person_1# **thinks that #Person_1# knows #Person_2# somewhere**, but #Person_2# denies it.
SUMMARY – D2: #Person_1# **thinks #Person_1# has met #Person_2#** somewhere, but #Person_2# thinks it's a mistake.
SUMMARY – D3: #Person_1# keeps asking where #Person_2#'s from because #Person_1# **thinks #Person_1# knows #Person_2#** but #Person_2# denies it.

UniLMv2: **#Person_1# finds out that #Person_2# used to work at common fitness gym and knows #Person_2# from somewhere else.** Then #Person_1# finds out that #Person_2# works at the whole bean café.

Transformer: #Person_2# doesn't know much about the test. #Person_1# is disappointed because it's troublesome. #Person_2# has no work at the cafe.

Commonsense Knowledge

DIALOGUE - E:
#Person_1#: Can't you come in and have tea now.
#Person_2#: Not yet, I must **water the garden** first.
#Person_1#: Do you have to water it now?
#Person_2#: I'm afraid I must. Look at it! It's terribly dry.
#Person_1#: What a nuisance!
#Person_2#: Last summer it was very dry, too. Don't you remember? I had to water it every day.
#Person_1#: Well, I'll have tea by myself. That was quick! Have you finished already?
#Person_2#: No, but look **out of the window.**
#Person_1#: I see, **it's raining.**
#Person_2#: That was a pleasant surprise. It means I can have tea, instead.

SUMMARY – E1: #Person_1# invites #Person_2# to have tea, but #Person_2# is busy watering the garden. **Then it rains, so #Person_2# stops and has tea with #Person_1#.**
SUMMARY – E2: #Person_1# invites #Person_2# to have tea, but #Person_2# is watering the garden. **Then it rains, #Person_2# can get in and have tea with #Person_1#.**
SUMMARY – E3: #Person_1# invites #Person_2# to have tea, but #Person_2# is watering the garden. **Then it rains, as a result, #Person_2# can get in and have tea with #Person_1#.**

UNILMV2: #Person_1# invites #Person_2# to have tea, **but #Person_2# has to water the garden first because it's raining.**

Transformer: #Person_1# 's willing to water #Person_2# and #Person_2# won't have tea .

Future Directions

- Constrain abstractive summarization to a more controllable range, with more reliable outputs
- Something beyond query
  - If you did not attend the meeting, how can you ask questions?
  - Queries are limited to templates, if you ask beyond templates, results tend to be bad
- More domain-specific knowledge
  - Meetings usually contain domain phrases/concepts

DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, Michael Zeng

DialSumm: A Real-Life Scenario Dialogue Summarization Dataset

Yulong Chen, Yang Liu, Liang Chen, Yue Zhang

QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization

Ming Zhong, Da Yin, Tao Yu, ...

MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization

Chenguang Zhu, Yang Liu, Jie Mei, Michael Zeng

Thanks!