Instructive Dialogue Summarization with Query Aggregations

Bin Wang, Zhengyuan Liu, Nancy F. Chen
I²R, A*STAR, Singapore

Institute for Infocomm Research
I²R

Task Definition

- Dialogue Summarization condenses dialogue information into shorter text. The output is conditioned on the input without considering user preferences.
- **Goal:** Summarize a dialogue with particular focus or aspect of interest by following instructions.
- Challenge: Lack of such training data

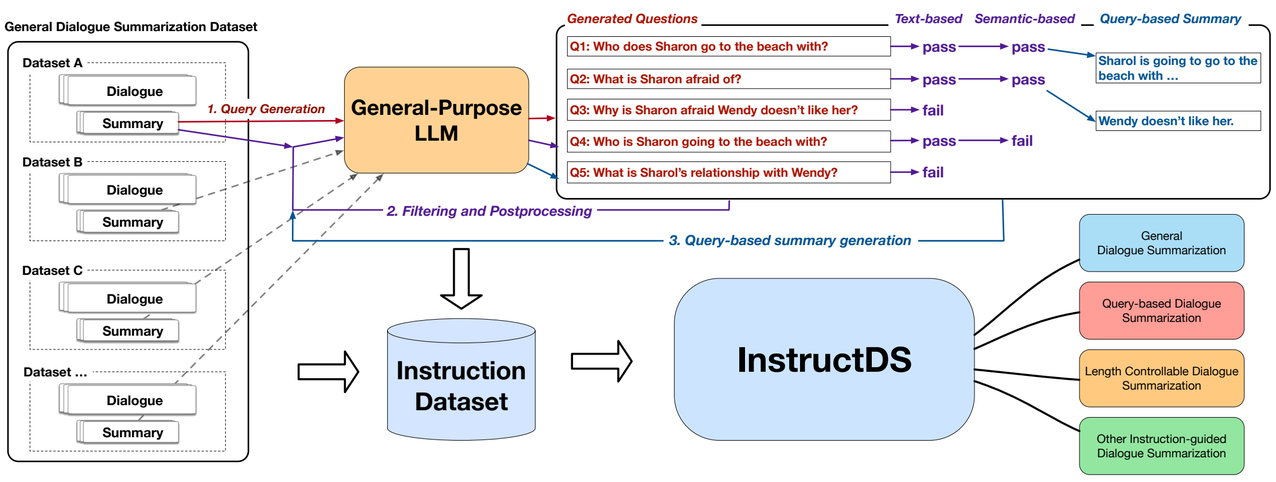
Benjamin: Hey guys, what are we doing with the keys today?
Hilary: I've got them. Whoever wants them can meet me at lunchtime or after
Elliot: I'm ok. We're meeting for the drinks in the evening anyway and
...
Benjamin: Interesting 😃 😱 To be honest, Hilary, I almost feel like changing my mind. Wanting to take this nap might end up costing me to dear
...
Hilary: Do join us, we're going to have fun. And then you'll take the keys and take this most deserved of naps
Elliot: Sounds like a plan
Hilary: 😃 😎
Elliot: See you at 2 then xx

Instruction: Where is Hilary having lunch?
Output: La Cantina

Instruction: Highlight the key takeaways from the dialogue.
Output: Hilary has the keys to the apartment. Benjamin wants to get them and go take a nap. Hilary is having lunch with some French people at La Cantina. Hilary ...

Instruction: Summarize the dialogue with about 15 words.
Output: Benjamin, Hilary and Elliot are discussing about their plans for the day after getting the apartment keys.

Methodology



Main contributions:
Summary-anchored-method to prompt LLM to synthesized Query-Dialogue-Summary triples
- Summaries are human-written condense information
- Covers important information
- Span multiple utterances

Model: Flan-T5-XL (3 billion parameters)

1. **Query Generation**
2. **Quality filtering**
   - Text-based filtering (Is the question answerable?)
   - Semantic-based filtering (Is the question unique?)
3. **Query-based Summary Generation**

Dialogue (SAMSum)
Emma: We are going beach would you like to join in? Sharol: sure who else is coming? Emma: its me anna, emily, wendy, kate and you Sharol: would wendy be ok with me joining.. Emma: i think she would be .. is there any problem between you guys? Sharol: i think she doesnt like me ... she always try to avoid me... Emma: really? then i think you should definetly join and sort things with her Sharol: hmm.. i dont want to be her friend forcefully... Emma: i know she doesnt dislike you there must b some misunderstanding... lets meet up and sort out.. be at my place at 11am Sharol: sure will be there
General Summary: Sharol is going to go to the beach with Emma, anna, emily, wendy and kate. Sharol is afraid that wendy doesn't like her.
Kept QDS Triples
Query: Who does Sharol go to the beach with? Summary: Sharol is going to go to the beach with Emma, Anna, Emily, Wendy and Kate.
Query: What is Sharol afraid of? Summary: wendy doesn't like her
Filtered QDS Triples
Query: Why is Sharol afraid Wendy doesn't like her? Reason for rejection: **text-based filtering**, the answer may not be answerable.
Query: Who is Sharol going to the beach with? Reason for rejection: **semantic-based filtering**, duplicate question.
Query: What is Sharol's relationship to Wendy? Reason for rejection: **text-based filtering**, the answer may not be answerable.

Quality Review Question	Yes% (w/o filtering)
Does the query question answerable?	94% (76%)
Is the query differs from previous ones for the same dialogue?	90% (63%)
Is the generated summary correct and acceptable for query and dialogue?	83% (71%)
Both unique and correct.	75% (45%)

Experimental Results

Datasets

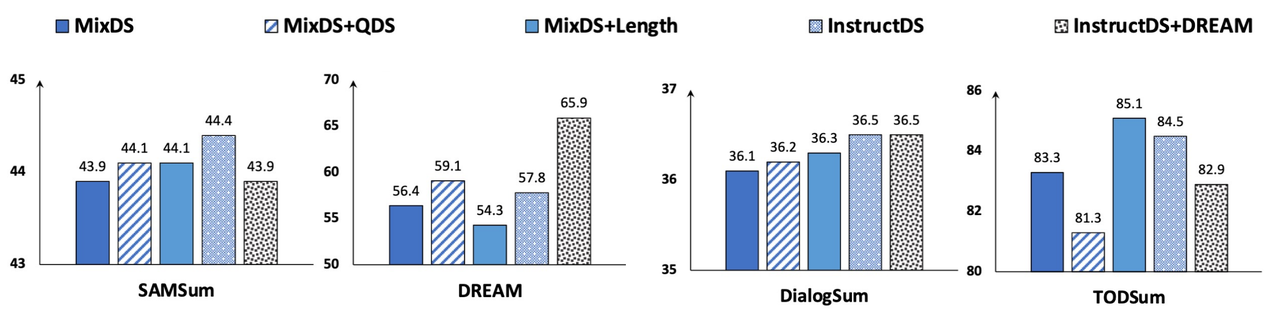
Dataset	# Train	# Validation	# Test	# QDS Triples
SAMSum (Gliwa et al., 2019)	14,732	818	819	18,245
DialogSum (Chen et al., 2021)	12,460	500	1,500	18,600
TODSum (Zhao et al., 2021)	7,892	999	999	8,705
DREAM (Sun et al., 2019)	6,116	2,040	2,041	-

- We train one unified model with all prepared data.

Results

Models	Params	ROUGE-1			ROUGE-2			ROUGE-L			BS
		F_1	*Pre*	*Rec*	F_1	*Pre*	*Rec*	F_1	*Pre*	*Rec*	
Pointer-Generator	-	40.1	-	-	15.3	-	-	36.6	-	-	-
BART	400M	53.0	59.0	52.8	28.4	32.1	28.2	44.2	49.3	44.0	53.3
MV-BART	400M	53.9	55.7	57.4	28.4	29.3	30.6	44.4	45.7	47.5	53.6
Coref-BART	400M	53.7	56.9	56.4	28.5	30.5	29.7	44.3	46.9	46.5	53.5
ConDigSum	400M	54.3	56.0	57.6	29.3	30.4	31.2	45.2	46.6	48.0	54.0
GPT-3-finetune	175B*	53.4	-	-	29.8	-	-	45.9	-	-	-
Alpaca	7B	28.2	26.0	39.8	5.7	5.1	8.3	20.5	19.2	29.0	19.4
Flan-T5-XXL	11B	52.6	62.6	50.0	28.5	34.1	27.1	44.1	52.5	41.9	53.2
Flan-UL2	20B	53.3	60.3	52.5	28.0	32.0	27.7	44.1	50.0	43.3	53.5
ChatGPT	175B	32.7	22.4	70.2	12.3	8.4	27.1	24.7	16.9	53.6	32.5
InstructDS	3B*	**55.3**	58.8	57.5	**31.3**	33.5	32.6	**46.7**	49.7	48.6	**55.5**
w/ reference summary length											
ChatGPT	175B	40.8	39.3	43.4	13.7	13.2	14.6	31.5	30.5	33.4	40.0
InstructDS	3B*	**58.4**	58.5	58.8	**32.8**	32.9	33.0	**48.9**	49.0	49.2	**58.5**

- Results on SAMSum. *other results can find in the paper



- Ablation Study on multiple components

Takeaway

Why better performance is obtained?
- Larger and specialized model
  - 3B model vs. 0.4B model (BART-Large)
  - One-model-for-all or InstructDS
- Query-Dialogue-Summary triples enforce the model to reason about the dialogue.
- Training with multiple dialogue summarization datasets within one unified model.

Our proposed data synthesized method can be applied to other fields.
- Use LLM for data synthesized with quality filtering steps.

Paper

GitHub

Bin Wang
Research Scientist
A*STAR I²R