NEURAL INFORMATION PROCESSING SYSTEMS

Topic-Conversation Relevance (TCR) Dataset and Benchmarks

Microsoft

Yaran Fan, Jamie Pool, Senja, Filipi, Ross Cutler

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Agenda

- Part 1. Introduction
- Part 2. Topic-Conversation Relevance (TCR) **Dataset**
  - Overview
  - Dataset Schema
  - Data Sources
  - Data Augmentation
- Part 3. Topic-Conversation Relevance **Benchmarks**
- Part 4. Future Work

Part 1. Introduction - Topic-Conversation Relevance

- **Context**
  - More **online** meetings
  - Lack of **focused discussions** in ineffective meetings
- **Goals**
  - To **help create tools** that behave as a virtual meeting moderator by keeping the discussion on-track.
  - Create a **dataset** and **benchmark** to **measure** **how relevant a conversation transcript is to a topic.**
- **ContributionS**
  1. We create a large topic-conversation **dataset** covering multiple domains of meetings. This dataset consists of the newly collected meetings and aggregated public data sources.
  2. We use GPT-4 to **rewrite** long and detailed meeting minutes into a pre-meeting agenda topic style.
  3. We provide a design of an extensible **schema** that allows users to create variations of meetings where topics can be flexibly added and removed.
  4. We open-source **scripts** for data augmentation and synthetic meeting creation on top of the TCR dataset.

Part 2. TCR Dataset

- **Original** data collection and **public** data sources
- Overall statistics (Table 1):
  - 1,506 unique meetings
  - 22 million words in transcripts
  - More than 15,000 meeting topics
- A balanced subset is available too (Table 2 in paper).

Table 1: Topic-Conversation Relevance (TCR) Dataset Statistics

Category	Data Name	Number of Meetings	Number of Topics	Words	Duration (Hours)
Unique Meetings	SIM	84	84	529,012	48.6
	SIM_syn100	100	348	500,825	45.7
	ICSI	75	489	767,437	70.4
	MeetingBank	1,100	6,595	19,626,469	2,493.8
	NCPC	20	160	423,305	47.2
	QMSum_AMI	96	510	489,961	54.4
	QMSum_Parliament	20	158	276,620	30.7
	ELITR	11	94	56,521	6.3
	Sub Total	**1,506**	**8,438**	**22,670,150**	**2,797**
Different Annotations	QMSUm_ICSI	52	288	527,206	48.8
	MeetingBank ReAnnotated	1,100	6,585	19,626,469	2,493.8

Part 2. TCR Dataset

Dataset Schema

- The data is in **JSON** format
- The meeting contents are grouped at **topic** level
- Easy to **add or remove** topics

- A simplified structure:

```
–Source
  –Meeting
       –Metadata
       –Topics
            –Topic info
            –Transcripts
                 –Text and info by lines
```

```
{
    "EXAMPLE_DATA_SOURCE": {  # Data Source Name
        "EXAMPLE_MEETING_NAME": { # Meeting Name
            "metadata": { # Metadata
                "topic_annotation_source": "EXAMPLE_ANNOTATION_SOURCE", # Annotation Source of Topics
                "timestamp_source": "EXAMPLE_TIMESTAMP_SOURCE", # Annotation Source of Timestamps
                "meeting_start_s": 0.0, # Start Time of Meeting in Seconds
                "meeting_end_s": 1800.0, # End Time of Meeting in Seconds
                "meeting_start_line": 0.0, # Start Line of Meeting in Transcripts
                "meeting_end_line": 200.0, # End Line of Meeting in Transcripts
                "meeting_trans_word_count": 3000.0, # Total Word Count of Meeting Transcripts
                "variations": {} # Type of variations included in the meeting.
                                 # Refer to 'script_augment_data.py' for creating meeting variations.
            },
            "topics": { # Topics of the Meeting
                "EXAMPLE TOPIC TEXT 1": { # Topic content
                    "topic_start_s": 0.0, # Start Time of Topic in Seconds
                    "topic_end_s": 500.16, # End Time of Topic in Seconds
                    "topic_start_line": 0.0, # Start Line of Topic in Transcripts
                    "topic_end_line": 77.0, # End Line of Topic in Transcripts
                    "topic_trans_word_count": 900.0, # Total Word Count of Topic Transcripts
                    "transcripts": [ # Transcripts of the Topic
                        {
                            "line_id": 0.0, # Line ID/Number
                            "speaker": "speaker_3", # Speaker Name
                            "start_s": 0.0, # Start Time of Transcript in Seconds
                            "end_s": 3.65, # End Time of Transcript in Seconds
                            "contents": "Hello everyone, welcome to today's talk.", # Transcript Content
                            "word_count": 6.0, # Word Count of Transcript
                            "cum_wc": 6.0 # Cumulative Word Count of Transcript
                        },
                        {
                            # Next Line of Transcript and Related Information with the same format
                        },
                        # ...
                    ]
                },
                "EXAMPLE TOPIC TEXT 2": {
                    # Next Topic and Transcripts with the same format
                }
            },
        }
    }
}
```

Part 2. TCR Dataset

Data Sources

❖ **New** Data: Speech Interruption Meetings (SIM)

- The SIM data captures natural online meeting dynamics.
- Data collection procedure:
  - Microsoft Teams (all participants are remote)
  - 4 participants per meeting
  - 149 unique speakers
  - One pre-defined topic for a 30 minutes session.
  - Natural interactions are strongly encouraged.
  - 84 meetings, 48 hours of data is included
- We also create 100 multi-topic synthetic meetings (SIM_syn100) on top of these raw meetings.

Table 3: Exploratory Analysis of Mean Metrics per Meeting

Data Source	Meeting Duration (minutes)	N Speakers per Meeting	N Topics per Meeting
SIM	34.24	4.00	1.00
SIM_syn100	27.24	4.00	3.48
ICSI	45.19	6.20	6.52
QMSum_ICSI	44.88	6.31	4.27
QMSum_AMI	34.03	4.00	3.94
QMSum_Parliament	92.21	23.80	6.45
MeetingBank	109.86	8.54	5.98
MeetingBank_ReAnnotated	109.86	8.54	5.97
NCPC	141.69	25.60	8.00
ELITR	34.26	5.45	7.64

❖ **Public** Data Sources

- ICSI
- QMSum (selected)
- MeetingBank (selected, reannotated)
- NCPC (selected)
- ELITR (selected)

Part 2. TCR Dataset

Data Augmentation

1. Create synthetic multi-topic discussions from SIM data: [script_create_synthetic_meetings_SIM.py]

2. Augment dataset by adding or removing topics: [script_augment_data.py]

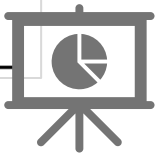
```
{
    "EXAMPLE_AUG_DATA": {  # Data Source Name
        "EXAMPLE_AUG_MEETING": { # Meeting Name
            "metadata": { # Metadata
                # Same metadata structure including data source, time and transcripts information
                # ...
                "variations": {
                    "variation_addToics": [ # Added topic list
                          "ADDED TOPIC A1 TEXT"],
                    "variation_removeTopics": [ # Removed topic list
                        "REMOVED TOPIC R1 TEXT"]
                }
            },
            "topics": { # Topics of the Meeting
                # Topic and Transcripts that are kept
                # ...
                "ADDED TOPIC A1 TEXT": {# Added planned topic
                    "topic_start_s": -1,
                    "topic_end_s": -1,
                    "topic_start_line": -1,
                    "topic_end_line": -1,
                    "topic_trans_word_count": 0,
                    "transcripts": []
                },
                "REMOVED TOPIC R1 TEXT": { # Remove topic and contents
                    # Contents of TOPIC R1
                },
            },
        }
    }
}
```

Part 3. Topic-Conversation Relevance Benchmarks

- **GPT-4**-32K
- Equal snippets (5m, 10m, 15m) of meetings.
- **Inputs**:
  - Snippet of **transcripts**
  - Full candidate **topic list**
- **Outputs**:
  - **Relevance class** for each topic
  - Classes:
    - 0 – Not Relevant
    - 1 – Somewhat Relevant
    - 2 – Mostly Relevant
    - 3 – Very Relevant
- **Metrics**:
  - **NOT discussed** class
  - **Precision**: the model says a topic is not discussed, and it indeed is the case
  - **Recall**: a topic is not discussed in the transcript, and the model picks it up

Table 4: Topic-Conversation Relevance Benchmark Results

Data Source	Transcripts Length	N Prompts	Prompt*Topic Pairs	F1	Precision	Recall
SIM_syn100	5 min	509	2,031	0.9587	0.9620	0.9554
	10 min	231	930	0.9272	0.8994	0.9568
	15 min	137	562	0.9175	0.8669	0.9744
ICSI_original75	5 min	790	5,175	0.8663	0.9615	0.7882
	10 min	382	2,502	0.8582	0.9462	0.7851
	15 min	244	1,594	0.8488	0.9243	0.7847
ICSI_QMSum	5 min	550	3,079	0.7506	0.9373	0.6259
	10 min	266	1,492	0.7242	0.9441	0.5874
	15 min	170	955	0.7222	0.9381	0.5871
MeetingBank _rnd30	5 min	594	4,236	0.9804	0.9891	0.9720
	10 min	301	2,168	0.9767	0.9843	0.9691
	15 min	204	1,479	0.9688	0.9671	0.9705
MeetingBank _ReAnnotated _rnd30	5 min	594	4,236	0.9817	0.9913	0.9723
	10 min	301	2,168	0.9810	0.9895	0.9726
	15 min	204	1,479	0.9755	0.9824	0.9687
NCPC	5 min	562	4,585	0.9702	0.9855	0.9553
	10 min	277	2,261	0.9631	0.9800	0.9468
	15 min	181	1,478	0.9614	0.9664	0.9565
ELITR	5 min	70	584	0.8429	0.9390	0.7646
	10 min	31	261	0.8182	0.9184	0.7377
	15 min	20	166	0.8043	0.8706	0.7475



Part 4. Future Work

- More types of meetings from **different domains**.
  - [*In-progress*] Collect more meetings by inviting domain experts (e.g., legal, healthcare, finance, etc.) to create meeting agendas for different types of meetings in their industry, and conducting domain-specific meetings based on the agendas.
- More **languages** other than English:
  - To translate the current data sources into other languages with reliable translation services and test the performance on the same tasks.
- A challenge of evaluating topic-conversation relevance is the blurred boundaries between topics
  - It would be desirable to create sub-labels at sentence or group of sentences level to capture relevance scores at a lower granularity.
- It would be beneficial to include **audio data** in the TCR dataset along with transcripts.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You

Project Repo: https://github.com/microsoft/topic_conversation

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.