Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization

Jiaao Chen
School of Interactive Computing
Georgia Institute of Technology
jiaaochen@gatech.edu

Diyi Yang
School of Interactive Computing
Georgia Institute of Technology
dyang888@gatech.edu

Abstract

Text summarization is one of the most challenging and interesting problems in NLP. Although much attention has been paid to summarizing structured text like news reports or encyclopedia articles, summarizing conversations—an essential part of human-human/machine interaction where most important pieces of information are scattered across various utterances of different speakers—remains relatively under-investigated. This work proposes a multi-view sequence-to-sequence model by first extracting conversational structures of unstructured daily chats from different views to represent conversations and then utilizing a multi-view decoder to incorporate different views to generate dialogue summaries. Experiments on a large-scale dialogue summarization corpus demonstrated that our methods significantly outperformed previous state-of-the-art models via both automatic evaluations and human judgment. We also discussed specific challenges that current approaches faced with this task. We have publicly released our code at https://github.com/GT-SALT/Multi-View-Seq2Seq.

1 Introduction

We live in an information age where communications between human and human/machine are increasing exponentially in the form of textual dialogues between users and users-agents (Kester, 2004). It is challenging and time-consuming to review all the content before starting any conversations especially when the chatting history becomes very long (Gao et al., 2020). How to process and organize those interaction activities into concise and structured data, i.e. conversation summarization, becomes technically and socially important.

Most existing research efforts on text summarization have been focused on single-speaker documents like news reports (Nallapati et al., 2016; See et al., 2017), scientific publications (Nikolov et al., 2018) or encyclopedia articles (Liu* et al., 2018), where structured text is usually used to elaborate a core idea in the third-person point of view, and the information flow is very clear through paragraphs or sections. Different from these structured documents, conversations are often informal, verbose and repetitive, sprinkled with false-starts, back channeling, reconfirmations, hesitations, speaker interruptions (Sacks et al., 1978) and the salient information is scattered in the whole chat, making current summarization models hard to focus on many informative utterances. Take the conversation in Table 1 as an example, turns, informal words, abbreviations, and emoticons all introduce new forms of challenges to the task of summarization. This calls for the design and development of new methods for dialogue summarization instead of directly applying current document summarization models.

There has been some recent research on conversation summarization such as directly deploying existing document summarization models (Gliwa et al., 2019) and exploring multi-sentence compression (Shang et al., 2018), however, most of them haven't utilized specific conversational structures, which refer to the way utterances are organized in order to make the conversation meaningful, enjoyable and understandable (Sacks et al., 1978), in dialogues – a key factor that differentiates dialogues from structured documents. As a way of using language socially of "doing things with words" together with other persons, the conversation has its own dynamic structures that organize utterances in certain orders to make the conversation meaningful, enjoyable, and understandable (Sacks et al., 1978). Although there are a few exceptions such as utilizing topic segmentation (Liu et al., 2019b; Li et al., 2019), dialogue acts (Goo and Chen, 2018) or key point sequence (Liu et al., 2019a), they either need

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4106–4118,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Conversation		**Topic View**	**Stage View**
James:	Hey! I have been thinking about you :)	Greetings	Openings
Hannah:	Oh, that's nice ;)		
James:	What are you up to?	Today's plan	
Hannah:	I'm about to sleep		Intention
James:	I miss u. I was hoping to see you		
Hannah:	Have to get up early for work tomorrow	Plan for tomorrow	Discussion
James:	What about tomorrow?		
Hannah:	To be honest I have plans for tomorrow evening		
James:	Oh ok. What about Sat then?	Plan for Saturday	
Hannah:	Yeah. Sure I am available on Sat		
James:	I'll pick you up at 8?	Pick up time	
Hannah:	Sounds good. See you then.		Conclusion
Summary	James misses Hannah. They agree for James to pick Hannah up on Saturday at 8.		

Table 1: Example conversation from SAMSum (Gliwa et al., 2019) with its topic view and stage view (extracted by our methods), and the human annotated summary.

extensive expert annotations of discourse acts(Goo and Chen, 2018; Liu et al., 2019a), or only encode conversations based on their topics (Liu et al., 2019b), which fails to capture rich conversation structures in dialogues.

Even one single conversation can be viewed from different perspectives, resulting in multiple conversational or discourse patterns. For instance, in Table 1, based on what topics were discussed (**topic view**) (Galley et al., 2003; Liu et al., 2019b; Li et al., 2019), it can be segmented into *greetings*, *today's plan*, *plan for tomorrow*, *plan for Saturday* and *pick up time*; from a conversation progression perspective (**stage view**) (Ritter et al., 2010; Paul, 2012; Althoff et al., 2016), the same dialogue can be categorized into *openings*, *intention*, *discussion*, and *conclusion*. From a coarse perspective (**global view**), conversations can be treated as a whole, or each utterance can serve as one segment (**discrete view**). Models that only utilized a fixed topic view of the conversation (Joty et al., 2010; Liu et al., 2019b) may fail to capture its comprehensive and nuanced conversational structures, and any amount of information loss introduced by the conversation encoder may lead to larger error cascade in the decoding stage. To fill these gaps, we propose to combine those multiple, diverse views of conversations in order to generate more precise summaries.

To sum up, our contributions are: (1) we propose to utilize rich conversational structures, i.e., structured views (topic view and stage view) and the generic views (global view and discrete view) for abstractive conversation summarization. (2) We design a multi-view sequence-to-sequence model that consists of a conversation encoder to encode different views and a multi-view decoder with multi-view attention to generate dialogue summaries. (3) We perform experiments on a large-scale conversation summarization dataset, SAMSum (Gliwa et al., 2019), and demonstrate the effectiveness of our proposed methods. (4) We conduct thorough error analyses and discuss specific challenges that current approaches faced with this task.

2 Related Work

Document Summarization Document summarization has received extensive research attention, especially for abstractive summarization. For instance, Rush et al. (2015) introduced to use sequence-to-sequence models for abstractive text summarization. See et al. (2017) proposed a pointer-generator network to allow copying words from the source text to handle the OOV issue and avoid generating repeated content. Paulus et al. (2018); Chen and Bansal (2018) further utilized reinforcement learning to select the correct content needed by summarization. Large-scale pre-trained language models (Liu and Lapata, 2019; Raffel et al., 2019; Lewis et al., 2019) have also been introduced to further improve the summarization performance. Other line of work explored long-document summarization by utilizing discourse structures in text (Cohan et al., 2018), introducing hierarchical models (Fabbri et al., 2019) or modifying attention mechanisms (Beltagy et al., 2020). There are also recent studies looking at the faithfulness in

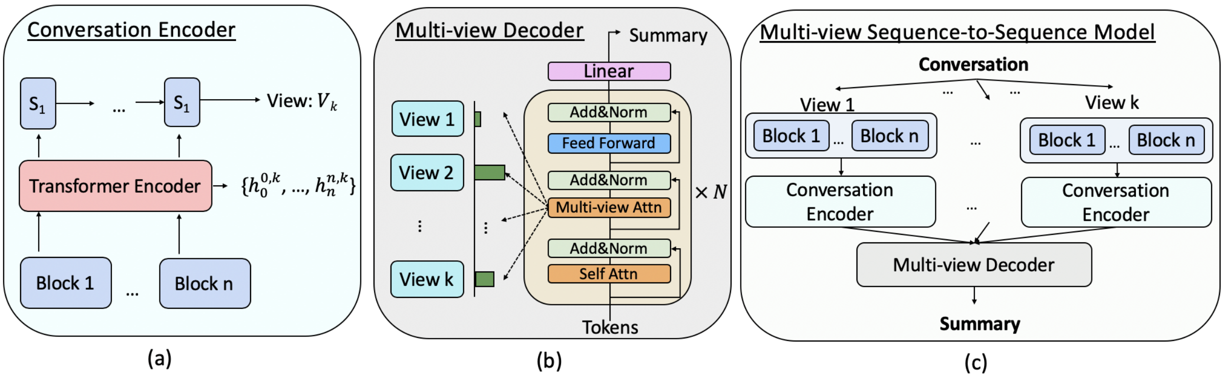

Figure 1: Model architecture. Different views of conversations are first extracted automatically, and then encoded through the conversation encoder (a) and combined in the multi-view decoder to generate summaries (b). In the conversation encoder, each view (consists of blocks) is encoded separately and the block's representations S_i are encoded through LSTM to represent the view. In the multi-view decoder, the model decides attention weights over different views and then attend to each token in different views through the multi-view attention.

document summarization (Cao et al., 2018; Zhu et al., 2020a), in order to enhance the information consistency between summaries and the input.

Dialogue Summarization When it comes to the summarization of dialogues, Shang et al. (2018) proposed a simple multi-sentence compression technique to summarize meetings. Zhao et al. (2019); Zhu et al. (2020b) introduced turn-based hierarchical models that encoded each turn of utterance first and then used the aggregated representation to generate summaries. A few studies have also paid attention to utilizing conversational analysis for generating dialogue summaries, such as leveraging dialogue acts (Goo and Chen, 2018), key point sequence (Liu et al., 2019a) or topics (Liu et al., 2019b; Li et al., 2019). However, they either needed a large amount of human annotation for dialogue acts, key points or visual focus (Goo and Chen, 2018; Liu et al., 2019a; Li et al., 2019), or only utilized topical information in conversations (Li et al., 2019; Liu et al., 2019b).

These prior work also largely ignored diverse conversational structures in dialogues, for instance, reply relations among participants (Mayfield et al., 2012; Zhu et al., 2019), dialogue acts (Ritter et al., 2010; Paul, 2012), and conversation stages (Althoff et al., 2016). Models that only utilized a fixed topic view of the conversation (Galley et al., 2003; Joty et al., 2010) may fail to capture its comprehensive and nuanced conversational structures, and any amount of information loss introduced by the conversation encoder may lead to larger error cascade in the decoding stage. To fill these gaps, we propose to leverage diverse conversational structures including topic segments, conversational stages, dialogue overview, and utterances to design a multi-view model for dialogue summarization.

3 Method

Conversations can be interpreted from different views and every single view enables the model to focus a specific aspect of the conversation. To take advantages of those rich conversation views, we design a Multi-view Sequence-to-Sequence Model (see Figure 1) that firstly extracts different views of conversations (Section 3.1) and then encodes them to generate summaries (Section 3.2).

3.1 Conversation View Extraction

Conversation summarization models may easily stray among all sorts of information across various speakers and utterances especially when conversations become long. Naturally, if informative structures in the form of small blocks can be explicitly extracted from long conversations, models may be able to understand them better in a more organized way. Thus, we first extract different views of structures from conversations.

Topic View Although conversations are often less structured than documents, they are mostly organized around topics in a coarse-grained structure (Honneth et al., 1988). For instance, a telephone chat could possess a pattern of "*greetings* → *invitation* → *party details* → *rejection*" from a topical perspective. Such explicit view and topic flow could help models interpret conversations more precisely and generate summaries that cover important topics. Here we combine the classic topic segment

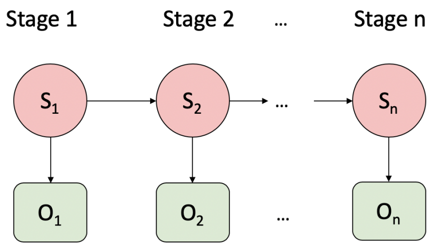

Figure 2: Allowed state transitions for the HMM conversation model. S_i are conversation stages, O_i are sentences' encoded representations. Conversation stages evolve in an increasing order from 1 to n.

Stage	Interpretation	Top Freq Words
1	Openings	hey, hi, good, yeah,going, time
2	Intentions	need, like, think, get, want, really
3	Discussions	will, know, time, come,tomorrow, meet
4	Conclusions	thanks, ok, see, great, thank, sure

Table 2: The top 6 frequent words appearing in each stage and the interpretations for different stages.

algorithm, C99 (Choi, 2000) that segments conversations based on inter-sentence similarities, with recent advanced sentence representations Sentence-BERT (Reimers and Gurevych, 2019), to extract the topic view. Specifically, each utterance u_i in a conversation $\mathbf{C} = \{u_1, u_2, ..., u_m\}$ is first encoded into hidden vectors via Sentence-BERT. Then the conversation $\mathbf{C}$ is divided into blocks $\mathbf{C}_{topic} = \{\mathbf{b}_1, ..., \mathbf{b}_n\}$ through C99, where $\mathbf{b}_i$ is one block that contains several consecutive utterances, such as the topic view described in Table 1.

Stage View As a way of doing things with words socially together with other people, conversation organizes utterances in certain orders to make it meaningful, enjoyable, and understandable. (Sacks et al., 1978; Althoff et al., 2016) For example, counseling conversations are found to follow a common pattern of "*introductions* → *problem exploration* → *problem solving* → *wrap up*" (Althoff et al., 2016). Such conversation stage view provides high-level sketches about the functions or goals of different parts in conversations, which could help models focus on the stages with key information.

We follow Althoff et al. (2016) to extract stages through a Hidden Markov Model (HMM). We impose a fixed ordering on the stages and only allow transitions from the current stage to the next one. The observations in the HMM model are the encoded representations h_i from Sentence-BERT. We set the number of hidden stages as 4. Similar to the topic view extraction, we segment the conversations into blocks $\mathbf{C}_{stage} = \{\mathbf{b}_1, ..., \mathbf{b}_n\}$, where $\mathbf{s}_i$ is one block that contains several consecutive utterances. We interpret the inferred stages qualitatively and further visualize the top 6 frequent words appearing in each stage in Table 2. We found that conversations around daily chats usually start with *openings*, introduce the goals/focus of the conversation followed by discussions of the details, and finally conclude with certain endings. Table 1 shows an example of the stage view.

Global View and Discrete View In addition to the aforementioned two structured views, conversations can also be naturally viewed from a relatively coarse perspective, i.e., a global view that concatenates all utterances into one giant block (Gliwa et al., 2019), and a discrete view that separates each utterance into a distinct block (Liu and Chen, 2019; Gliwa et al., 2019).

3.2 Multi-view Sequence-to-Sequence Model

We extend generic sequence-to-sequence models to encode and combine different conversation views. To better utilize semantic information in recent pre-trained models, we implement our base encoders and decoders with a transformer based pre-trained model, BART (Lewis et al., 2019). Note that our multi-view sequence-to-sequence model is agnostic to BART with which it is initialized.

Conversation Encoder Given a conversation under a specific view k with n blocks: $\mathbf{C}_k = \{\mathbf{b}_1^k, ..., \mathbf{b}_n^k\}$, each token $x_{i,j}^k$ in a block $\mathbf{b}_j^k = \{x_{0,j}^k, x_{1,j}^k, ..., x_{m,j}^k\}$ is first encoded through the conversation encoder $\mathbf{E}$, e.g., BART encoder as shown in Figure 1(a), into hidden representations:

$$\{h_{0,j}^k, h_{1,j}^k, ..., h_{m,j}^k\} = \mathbf{E}(\{x_{0,j}^k, x_{1,j}^k, ..., x_{m,j}^k\}) \tag{1}$$

Note that we add special tokens $x_{0,j}^k$ at the beginning of each block and use these tokens' representations to describe each block, i.e., $S_j^k = h_{0,j}^k$.

To depict different views using hidden vectors, we aggregate the information from all blocks in one conversation through LSTM layers (Hochreiter and Schmidhuber, 1997):

$$S_j^k = \text{LSTM}(h_0^{j,k}, S_{j-1}^k), j \in [1, n] \tag{2}$$

We use the last hidden state S_n^k to represent the current view k, denoted as V_k.

Multi-view Decoder Different views could provide different types of conversational aspects for models to learn and further determine which set of utterances should deserve more attention in order to generate better dialogue summaries. As a result, the ability to strategically combine different views is essential. To this end, we propose a transformer based multi-view decoder to integrate encoded representations from different views and generate summaries as shown in Figure 1(b).

The input to the decoder contains $l-1$ previously generated tokens $t_1, ..., t_{l-1}$. Via our multi-view decoder $\mathbf{D}$, the l-th token is predicted via:

$$\{y_1, ..., y_{l-1}\} = \mathbf{D}(\{t_1, ..., t_{l-1}\}, \mathbf{E}(C)) \quad (3)$$

$$P(\tilde{t}_l | t_{<l}, \mathbf{C}) = \text{Softmax}(W_p y_{l-1}) \quad (4)$$

Here, W_p is a parameter to be learned.

Different from generic transformer decoder, we introduce a multi-view attention layer in each transformer block. Multi-view attention layer first decides the importance α_k of each view V_k through:

$$u_k = \tanh\left(W V_k + b\right) \quad (5)$$

$$\alpha_k = \frac{\exp\left(u_k^\top v\right)}{\sum_i \exp\left(u_i^\top v\right)} \quad (6)$$

where v is a randomly initialized context vector; W and b are parameters. To avoid the attention weights being too similar to each other as views are actually encoded from a similar context, we utilize a sharpening function over α_k with a temperature T: $\tilde{\alpha}_k = \alpha_k^{\frac{1}{T}} / \sum_i \alpha_i^{\frac{1}{T}}$. When $T \rightarrow 0$, the attention weights will behave like a one-hot vector.

Then the multi-head attention is performed over conversation tokens $h_{i,j}^k$ from different views k and form A^k separately. The attended results are further combined based on the view-attention weights $\tilde{\alpha}_k$ and continue forward passing:

$$\tilde{A} = \sum_k \tilde{\alpha}_k A^k \quad (7)$$

Training We minimize the cross entropy loss during training:

$$L = -\sum \log P(\tilde{t}_l | t_{<l}, \mathbf{C}) \quad (8)$$

Specifically, we apply the teacher forcing strategy: at training time, the inputs are previous tokens from the ground truth; at test time, the inputs are previous tokens predicted by the decoder.

4 Experiments

4.1 Dataset and Baselines

We evaluate our model on a large-scale dialogue summary dataset SAMSum (Gliwa et al., 2019) that has 14732 dialogues with human-written summaries. The data statistics are shown in Table 3. SAMSum contains messenger-like conversations about daily topics, such as chit-chats, arranging meetings, discussing events, etc. We compare our Multi-view Sequence-to-Sequence Model (Multi-view BART) with several baseline models:

- **Pointer Generator** (See et al., 2017): Following Gliwa et al. (2019), we added separators between each utterance (**discrete view**) and used it as input for pointer generator model.

- **DynamicConv + GPT-2/News** (Wu et al., 2019): We followed Gliwa et al. (2019) to use GPT-2 to initialize token embeddings (Radford et al., 2019). We also added news summarization corpus CNN/DM (Nallapati et al., 2016) as extra training data.

- **Fast Abs RL Enhanced** (Chen and Bansal, 2018) first selects salient sentences and then rewrites them abstractively via sentence-level policy gradient methods. We combined it with the **global view** (Gliwa et al., 2019).

- **BART + Generic views** (Lewis et al., 2019) utilized BART, a denoising autoencoder for pretraining sequence-to-sequence models, together with generic views (**global view** and **discrete view**). We used the BART-large model with its default settings [1].

4.2 Model Settings[2]

We loaded the pre-trained "bert-base-nli-stsb-mean-tokens"[3] for sentence-BERT to get representations for each utterance. For extracting the *topic view* via C99, we set the window size 4 and std coefficient 1. For extracting the *stage view*, we set the number of hidden states 4 in HMM. These hyper-parameters were set with a grid search. The **BART + Structured views** (stage and topic views) used the same set of parameters as **BART + Generic views**. For

[1] https://github.com/pytorch/fairseq

[2] More details are shown in Section A in the Appendix.

[3] https://github.com/UKPLab/sentence-transformers

# Conversations		# Participants			# Turns			Reference Length		
		Mean	Std	Interval	Mean	Std	Interval	Mean	Std	Interval
Train	14732	2.40	0.83	[1, 14]	11.17	6.45	[1, 46]	23.44	12.72	[2, 73]
Dev	818	2.39	0.84	[2, 12]	10.83	6.37	[3, 30]	23.42	12.71	[4, 68]
Test	819	2.36	0.83	[2, 11]	11.25	6.35	[3, 30]	23.12	12.20	[4, 71]

Table 3: SAMSum dataset statistics. *Interval* denotes the minimum and maximum range.

Model	Views	ROUGE-1			ROUGE-2			ROUGE-L		
		F	P	R	F	P	R	F	P	R
Pointer Generator	Discrete	0.401	-	-	0.153	-	-	0.366	-	-
DynamicConv + GPT-2	Global	0.418	-	-	0.164	-	-	0.376	-	-
Fast Abs RL Enhanced	Global	0.420	-	-	0.181	-	-	0.392	-	-
DynamicConv + News	Discrete	0.454	-	-	0.206	-	-	0.415	-	-
BART	Discrete	0.481	0.452	0.526	0.245	0.236	0.282	0.451	0.432	0.521
	Global	0.482	0.493	0.517	0.245	0.251	0.264	0.466	0.475	0.495
BART†	Stage	0.487	0.483	0.540	0.251	0.248	0.282	0.472	0.469	0.515
	Topic	0.488	0.479	0.547	0.251	0.248	0.284	0.474	0.483	0.501
Multi-view BART†	Global + Stage	0.488	0.476	0.548	0.251	0.246	0.285	0.472	0.462	0.521
	Global + Topic	0.488	0.488	0.535	0.251	0.252	0.275	0.473	0.474	0.509
	Topic + Stage	**0.493**	0.511	0.522	**0.256**	0.265	0.274	**0.477**	0.493	0.499

Table 4: ROUGE-1, ROUGE-2 and ROUGE-L scores for different models on the test set. Results are averaged over three runs. † meant our methods or utilized views introduced by us.

Multi-View BART, we experimented with different view combinations: (1) the best generic view - global view, was combined with two structured views (stage and topic view) separately; (2) the best two structured views are also combined (topic + stage). The settings for BART encoder/decoder kept identical as baselines. We used a one-layer LSTM for encoding sections. The learning rate for section encoder and multi-view attention was set 3e-3. The temperature T was 0.2. The beam search size during inference for all the models was 4.

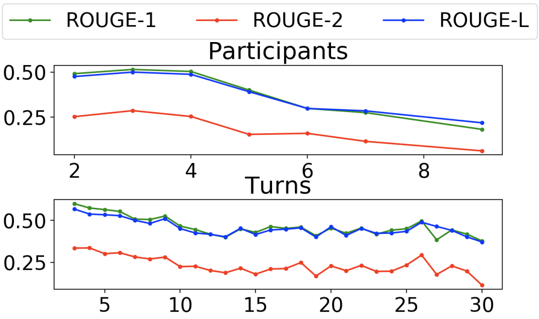

Figure 3: Relations between ROUGE scores and the number of participants/turns in conversations.

4.3 Results

Quantitative Results We evaluated models with the standard metric ROUGE Score (with stemming) (Lin and Och, 2004), and reported ROUGE-1, ROUGE-2 and ROUGE-L[4]. Results on the test set for different models were shown in Table 4. Compared to *Pointer Generator*, using reinforcement learning to select important sentences first (*Fast Abs RL Enhanced*) slightly increased F scores. Adding pre-trained embeddings or extra documents training data to lightweight convolution models, (*DynamicConv + GPT-2/News*) lead to even better ROUGE scores. When using pre-trained transformer based model BART with generic views, all ROUGE scores improved significantly, and *BART + Global* outperformed *BART + Discrete* especially in terms of ROUGE-L F scores. Segmenting conversations into blocks from structured views (stage view and topic view) further boosted the performance, suggesting that our extracted conversation structures help conversational encoders to capture nuanced and informative aspects of dialogs.

We did not see any performance boost when combining the generic global view with either topic or conversational stage views, partially due to that the coarse granularity of global view does not complement structured views well. In contrast, utilizing both structured views (topic view + stage view) further increased ROUGE scores consistently, indicating the effectiveness of synthesizing informative conversation blocks introduced by both views.

We visualized the attention weight distributions

[4] Here we followed BART and used `https://github.com/pltrdy/rouge`. Note that different tools may generate different ROUGE scores.

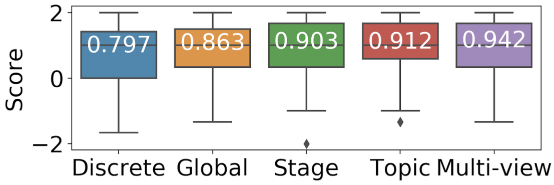

Figure 4: Human evaluation results. The mean score for each model is also shown in the box plot.

for the stage view and topic view in our best model (see Appendix) and found contributions of topic views are slightly more prominent compared to stage views. This also communicated that the two different structured views can complement each other well though sharing the same dialogue content. Note that the gains from *Multi-view BART (Topic + Stage)* are mainly from the precision scores while recall scores are kept comparable, suggesting that our proposed model produced fewer irrelevant tokens while preserving necessary information in its generated summary.

Impact of Participants and Turns We visualized the impact of two essential components in conversations—the number of participants and turns—on rouge scores via our best-performing model *Multi-view BART with topic view + stage view* in Figure 3. As the number of participants/turns increases, ROUGE scores decrease, indicating that the difficulty of conversation summarization increased with more participants involved in conversations and more utterances.

Qualitative/Human Evaluation We also conducted human annotations to evaluate the extracted dialogue summaries, in addition to ROUGE scores. Similar to Gliwa et al. (2019), we asked human annotators on Amazon Mechanical Turk [5] to rate each summary (200 randomly sampled summaries in total) on the scale of [-2, 0, 2], where -2 means that a summary was poor, extracted irrelevant information or did not make sense at all, 2 means it was understandable and gave a concise overview of the text, and 0 refers to that the summary only extracted only a part of relevant information, or made some mistakes. The score for each summary was averaged among three different annotators. The Intra-class Correlation was 0.583, indicating moderate agreement (Koo and Li, 2016).

As shown in Figure 4, consistent with ROUGE scores in Table 4, our multi-view model achieved the highest human annotation scores, significantly higher (via a student t-test) than either generic (discrete or global) view or structured (stage or topic) view, which further proved the effectiveness of combing different views.

5 Model Analysis and Discussion

So far, we have achieved a reasonable summarization performance. To further study why dialog summarization is challenging and how future research could advance this direction, we take a closer look at this dialogue summarization dataset (SAMSum), model generation errors, as well as certain challenges that existing approaches are struggling with.

5.1 Challenges in Dialog Summarization

We conduct a thorough examination of the challenges in conversation summarization and organized them into 7 categories as below:

1. **Informal language use** Many conversations especially in online contexts such as Twitter/Reddit (Jackson and Moulinier, 2007), contain typos, word abbreviations, slang or emoticons/emojis, making it hard to be represented and summarized.

2. **Multiple participants** As shown in Figure 3, conversations with more speakers are harder to be summarized since it may require models to accurately differentiate both language styles and content from different speakers, similar to the multiple characters issue in story summarization (Zhang et al., 2019).

3. **Multiple turns** Similar to long document summarization (Xiao and Carenini, 2019), conversations with many utterances contain more information to be processed, thus harder to be summarized.

4. **(Referral and coreference** People usually refer to each other, mention others' names or use coreference in their messages, which introduces extra difficulty to dialogue summarization, also a challenge also exists in reading comprehension (Chen et al., 2016) and document summarization (Falke et al., 2017).

5. **Repetition and interruption** Information is generally scattered through the whole conversation, and speakers may interrupt each other,

[5] https://www.mturk.com/

Challenge	**%**	**ROUGE-1/2/L**
Generic	24	**0.613 / 0.384 / 0.579**
Informal language	25	0.471 / 0.241 / 0.459
Multiple participants	10	0.473 / 0.243 / 0.461
Multiple turns	23	0.432 / 0.213 / 0.432
Referral & coreference	33	0.445 / 0.206 / 0.430
Repetition & interruption	18	0.423 / 0.180 / 0.415
Negations & rhetorical	20	0.458 / 0.227 / 0.431
Role & language change	30	0.469 / 0.211 / 0.450

Table 5: The breakdown of challenges in dialogue summarization based on our analyses of 100 sampled conversations, and the ROUGE scores per challenge

reconfirm, back channeling or repeat themselves, a unique discourse challenge for dialogue summarization.

6. **Negations and rhetorical questions** As a long-standing problem in NLP field (Li et al., 2016), negation related issues are even more frequent in conversations, as there are more question-answer exchanges between speakers.

7. **Role and language change** Conversations usually involve more than one speaker, and the role of a speaker may shift from a questioner to an answerer, requiring the summarization model to dynamically deal with speaker roles and the associated language (e.g., first personal pronouns)

We randomly sampled 100 examples[6] from our test set and classified them using the above challenge taxonomy. A conversation might have more than one category labels, and if it had none of the aforementioned challenges, we labeled it as **(0) Generic**. Usually, the one marked as *Generic* were shorter or had a simple structure.

Table 5 presents the percentage of each type of challenge and per-category performances from our best model (*Multi-view BART with Topic view + Stage view*). We observed that: (i) *Referral & coreference* (33%) and *Role & language change* (30%) were the two most frequent challenges that dialogue summarization task faced. (2) As expected, *Generic* conversations were relatively easier summarize. (3) Our best model performed relatively worse when it came to *Repetition & interruption*, *Multiple turns*, and *Referral & coreference*, calling for more intelligent summarization methods to tackle those challenges.

Errors	**%**	**ROUGE-1/2/L**
Other	24	**0.611 / 0.363 / 0.584**
Missing information	37	0.448 / 0.236 / 0.445
Redundancy	13	0.442 / 0.231 / 0.441
Wrong references	27	0.460 / 0.232 / 0.454
Incorrect reasoning	24	0.447 /0.187 / 0.411
Improper gendered pronouns	6	0.421 / 0.212 / 0.428

Table 6: The common error types of our model compared to golden reference on 100 sampled conversations, and the ROUGE scores per error type.

5.2 Error Analysis[7]

We examined summaries generated by our best-performing model compared to ground-truth summaries, and observed several major error types:

1. **Missing information**: content mentioned in references is missing in generated summaries.

2. **Redundancy**: content occurred in generated summaries was not mentioned by references.

3. **Wrong references**: generated summaries contain information that is not faithful to the original dialogue, and associate one's actions/locations with a wrong speaker.

4. **Incorrect reasoning**: generated summaries reasoned relations in dialogues incorrectly, thus came to wrong conclusions.

5. **Improper gendered pronouns**: summaries used improper gendered pronouns (e.g., the misuse of gendered pronouns).

We annotated the same set of 100 randomly sampled summaries via the above error type taxonomy. A summary might have more than one category labels and we categorized a summary as **(0) Other** if it did not belong to any error types.

Table 6 presents the breakdown of error types and per-category ROUGE scores. We found that: (i) *missing information* (37%) was the most frequent error type, indicating that current summarization models struggled with identifying key information. (ii) *Incorrect reasoning* had a percentage of 24% with the worst ROUGE-2; despite of being a minor type 6%, *improper gendered pronouns* seemed to severely decrease both ROUGE-1 and ROUGE-2. (iii) The relatively low ROUGE scores associated with *incorrect reasoning* and *wrong references* urged better summarization models in dealing with faithfulness in dialogue summarization.

[6] The full analyzed set of examples are shown in Appendix.

[7] Error analysis for baselines are displayed in the Appendix.

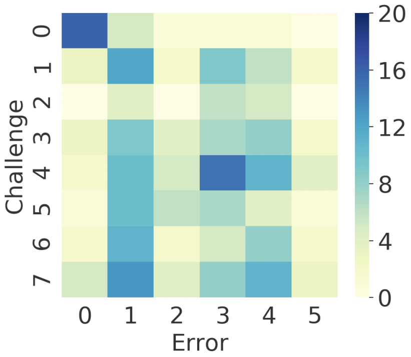

Figure 5: Relations between difficulties in conversations and errors made by our model.

5.3 Relation between Challenges and Errors

To figure out relations between challenges and errors made by our models, i.e., how different types of errors correlate with different types of challenges, we visualized the co-occurrence heat map in Figure 5. We found that: (i) Our model generated good summary for *generic*, simple conversations. (ii) All kinds of challenges had high correlations with, or could lead to the *missing information* error. (iii) *Wrong references* were highly associated with *referral & coreference*; this was as expected since co-references in conversations would naturally increase the difficulty for models to associate correct speakers with correct actions. (iv) High correlations between *role & language change*, *referral & coreference* and *incorrect reasoning* indicated that interactions between multiple participants with frequent co-references might easily lead current summarization models to reason incorrectly.

6 Conclusion

In this work, we proposed a multi-view sequence-to-sequence model that leveraged multiple conversational structures (topic view and stage view) and generic views (global view and discrete view) to generate summaries for conversations. In order to strategically combine these different views for better summary generations, we propose a multi-view sequence-to-sequence model. Experiments conducted demonstrated the effectiveness of our proposed models in terms of both quantitative and qualitative evaluations. Via thorough error analyses, we concluded a set of challenges that current models struggled with, which can further facilitate future research on conversation summarization. Due to the lack of annotations, we only adopted simple unsupervised segmentation methods to extract different views. In the future, we plan to annotate some of the data, explore supervised segmentation models (Li et al., 2018) and introduce more conversation structures like dialogue acts (Oya and Carenini, 2014; Joty and Hoque, 2016) into abstractive dialogue summarization.

Acknowledgment

We would like to thank the anonymous reviewers for their helpful comments, and the members of Georgia Tech SALT group for their feedback. We acknowledge the support of NVIDIA Corporation with the donation of GPU used for this research.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *AAAI*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tobias Falke, Christian M Meyer, and Iryna Gurevych. 2017. Concept-map-based multi-document summarization using concept coreference resolution and global importance optimization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 801–811.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2020. From standard summarization to new tasks and beyond: Summarization with manifold information.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *2018 IEEE Spoken Language Technology Workshop (SLT)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Axel Honneth, Hans Joas, et al. 1988. *Social action and human nature*. CUP Archive.

Peter Jackson and Isabelle Moulinier. 2007. *Natural language processing for online applications: Text retrieval, extraction and categorization*, volume 5. John Benjamins Publishing.

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond T. Ng. 2010. Exploiting conversation structure in unsupervised topic segmentation for emails. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 388–398, Cambridge, MA. Association for Computational Linguistics.

Shafiq Joty and Enamul Hoque. 2016. Speech act modeling of written asynchronous conversations with task-specific embeddings and conditional structured models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1756.

Grant H Kester. 2004. *Conversation pieces: Community and communication in modern art*. Univ of California Press.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Jing Li, Aixin Sun, and Shafiq R Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *IJCAI*, pages 4166–4172.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD19, page 1957–1965, New York, NY, USA. Association for Computing Machinery.

Peter J. Liu*, Mohammad Saleh*, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Zhengyuan Liu and Nancy Chen. 2019. Reading turn by turn: Hierarchical attention architecture for spoken dialogue comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5460–5466, Florence, Italy. Association for Computational Linguistics.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019b. Topic-aware pointer-generator networks for summarizing spoken conversations. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2012. Hierarchical conversation structure prediction in multi-party chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 60–69, Seoul, South Korea. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Nikola I. Nikolov, Michael Pfeiffer, and Richard H. R. Hahnloser. 2018. Data-driven summarization of scientific articles. *CoRR*, abs/1804.08875.

Tatsuro Oya and Giuseppe Carenini. 2014. Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 133–140.

Michael J. Paul. 2012. Mixed membership Markov models for unsupervised conversation modeling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 94–104, Jeju Island, Korea. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *EMNLP/IJCNLP*.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating character descriptions for automatic summarization of fiction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7476–7483.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, WWW '19, page 3455–3461, New

York, NY, USA. Association for Computing Machinery.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020a. Boosting factual correctness of abstractive summarization.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020b. End-to-end abstractive summarization for meetings.

Henghui Zhu, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Who did they respond to? conversation structure modeling using masked hierarchical transformer.

A Model Settings

We load the pre-trained "bert-base-nli-stsb-mean-tokens"[8] for sentence-BERT to get representations for each utterance. When extracting the **topic view**, we set the window size 4 and std coefficient 1 in C99. When extracting the **stage view**, we set the number of hidden states 4 in HMM. These hyper-parameters were set after a grid search with evaluating randomly sampled segmented results by human. The **BART + Structured views** (stage and topic views) followed the same parameters as **BART + Generic views**.

For **Multi-View BART**, we selected different views to combine: (1) generic view + structured view: best generic view, global view, was combined with two structured views (stage and topic view); (2) structured view + structured view: best two single views are combined (topic + stage). The settings for BART encoder/decoder kept the same as baseline. We used a one layer LSTM for encoding sections. The learning rate for section encoder and multi-view attention was set 3e-3. The temperature T was 0.2. The beam search size during inference for all the models was 4.

Experiments were performed on two Tesla P100 (16GB memory).

B View Attention Visualization

We visualized the attention weights distribution for the stage view and topic view in our best multi-view model to explore the importance of stage verses topic in Figure 6.We found that the topic views were more prominent than the stage views, consistent with the performances of *BART + topic view* and *BART + stage view*. This indicated that having discourse structures about topics might be more important while both topic and stage could improve the conversation summarization. This also communicated that the two different structured views can complement each other well though sharing the same dialogue content.

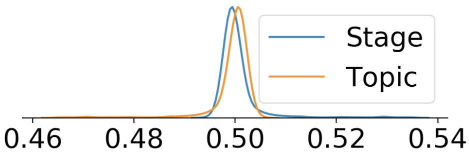

Figure 6: Attention weights distribution for stage view and topic view in the multi-view model.

330	191	635	733	342
595	454	629	598	466
158	576	676	344	353
621	255	106	66	742
446	327	497	463	478
320	258	528	405	305
208	550	512	663	165
69	431	796	338	443
254	716	549	51	145
364	259	190	479	182
617	189	422	177	8
741	151	488	176	212
15	124	461	386	197
172	372	508	323	162
793	308	486	763	376
493	520	116	513	802
358	784	53	655	23
717	374	289	64	217
519	539	441	341	350
136	713	426	648	355

Table 7: A full index list of our samples.

We displayed two examples in Table 8 with the golden references, each single view's generated summaries and the combined views' generated summaries. The combined view could balance the advantages of each single view and generated more precise summaries. And the attention weights the model learned were also consistent with single view's performances.

C Supplementary Examples for Model Analysis and Discussion

For the analysis in the **Model Analysis and Discussion** section in our paper, we randomly sampled 100 examples from the test set of the SAMSum

[8] https://github.com/UKPLab/sentence-transformers

Reference	James misses Hannah. They agree for James to pick Hannah up on Saturday at 8.	Petra is very sleepy at work today, Andy finds the day boring, and Ezgi is working.
Stage	Hannah has to get up early for wo--rk tomorrow. James will pick her up at 8 on Saturday. [0.61/0.13/0.40]	Petra needs to sleep, because she's sleepy. Ezgi is working. [0.37/0.16/0.38]
Topic	James and Hannah will see each other on Saturday at 8. [0.46/0.25/0.50]	Nobody is working at the office today. Ezgi is working. Petra is sleepy and wants to sleep. [0.53/0.19/0.53]
Stage + Topic	James will pick Hannah up on Saturday at 8 pm. [0.64/0.52/0.69]	Petra is sleepy and needs to sleep. Ezgi is working at the office. [0.60/0.21/0.43]
Attention Weight	[0.52, 0.48]	[0.45. 0.55]

Table 8: Some generated summary examples compared to references. [Rouge-1/Rouge-2/Rouge-L] is shown after each summary, and [stage weight/topic weight] is displayed in the last row.

Errors	**Discrete**	**Global**	**Stage**	**Topic**	**Multi-view**
Other	16	19	21	22	24
Missing information	40	46	45	42	37
Redundancy	33	44	18	25	13
Wrong references	32	33	26	30	27
Incorrect reasoning	27	28	22	28	24
Improper gendered pronouns	5	6	6	6	6

Table 9: Common error types of different models compared to golden reference on 100 sampled conversations.

dataset which can be downloaded here [9]. Table 7 provides a full index list of the samples.

Table 9 shows the error analysis for *BART-Discrete, BART-Global, BART-Stage, BART-Topic* and *BART-Multi-view* models. It can be observed that, (i) without any explicit structures, discrete-view and global-view models generated summaries with more *redundancies* compared to golden reference summaries, as models may easily lost focus on massive information; (ii) once we introduced certain conversation structures such as topic-view and stage-view, models behaved better in terms of *redundancy* and *incorrect reasoning*, which indicated that the structured views could help models to better understand the conversations; (iii) our multi-view models which combined both stage-view and topic-view made the least number of errors compared to all single view models, suggesting the effectiveness of combining different views for conversation summarization.

[9] https://arxiv.org/abs/1911.12237