Narrate Dialogues for Better Summarization

Ruochen Xu, Chenguang Zhu, Michael Zeng
Azure Cognitive Services Research, Microsoft
`{ruox, chezhu, nzeng}@microsoft.com`

Abstract

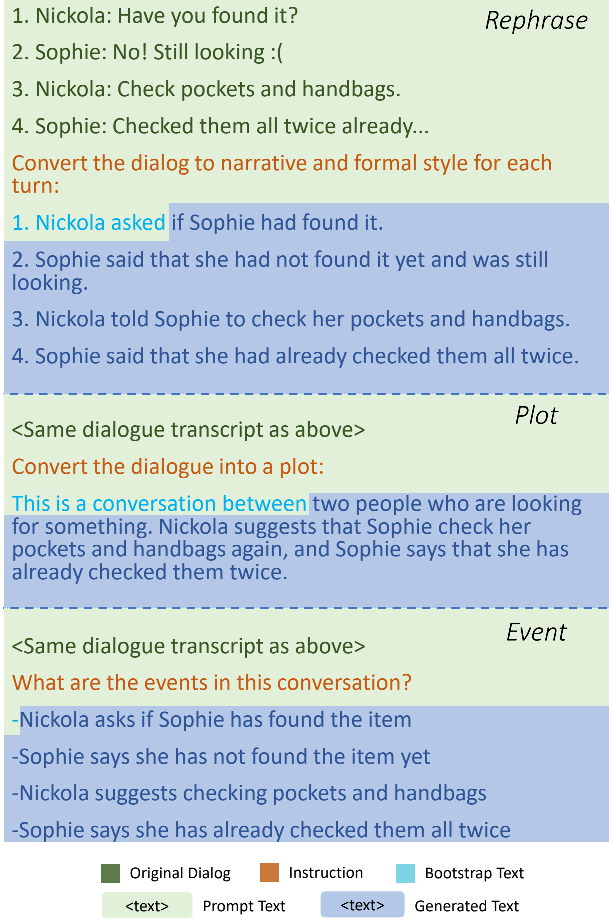
Dialogue summarization models aim to generate a concise and accurate summary for multi-party dialogue. The complexity of dialogue, including coreference, dialogue acts, and inter-speaker interactions bring unique challenges to dialogue summarization. Most recent neural models achieve state-of-art performance following the pretrain-then-finetune recipe, where the large-scale language model (LLM) is pretrained on large-scale single-speaker written text, but later finetuned on multi-speaker dialogue text. To mitigate the gap between pretraining and finetuning, we propose several approaches to convert the dialogue into a third-person narrative style and show that the narration serves as a valuable annotation for LLMs. Empirical results on three benchmark datasets show our simple approach achieves higher scores on the ROUGE and a factual correctness metric.

1 Introduction

Online dialogues are increasingly important in the modern working environment, emphasizing the need for an automatic system to generate concise and accurate summaries. Neural dialogue summarization has become an emerging research direction in recent years (Feng et al., 2021a) with the creation of several benchmarks (Gliwa et al., 2019; Mehnaz et al., 2021; Fabbri et al., 2021; Zhong et al., 2021; Zhu et al., 2021a). Most works utilized large-scale language models (LLM) and finetune it on downstream dialogue summarization datasets. Despite the strong generalization power of LLM on summarization tasks (Lewis et al., 2020; Zhang et al., 2020a), dialogue summarization holds some unique challenges (Feng et al., 2021a). First, dialogues and their summaries are in different language styles. This discrepancy requires the summarization model to complete the tasks of both style transfer and summarization. The domain discrepancy also exists between the pretraining and finetuning stages because LLMs are often pretrained on the web corpus where the majority of text is in written language. The dialogues, on the other hand, are in spoken language. Second, the amount of training data is generally smaller than news summarization. For instance, the widely used SAMSum dataset (Gliwa et al., 2019) for dialogue summarization contains about 16k annotations, while the CNN/DailyMail dataset (Nallapati et al., 2016) is a magnitude larger at size of 300k. Third, dialogues contain complex dialogue acts with frequent topic changes and event occurrences. News or academic articles follow certain patterns in writing and could be exploited for summarization (Zhu et al., 2021b). On the other hand, key information is scattered in the dialogues. To generate comprehensive summaries, a model needs to identify salient information across the dialogue and rephrase the terms.

To address these challenges, existing works incorporate external models or tools to help with the dialogue summarization (Liu and Chen, 2021; Feng et al., 2021b; Wu et al., 2021; Liu et al., 2021b). However, they either annotate the dialogues on the token or utterance level, such as coreference resolution (Liu et al., 2021b), personal named entity (Liu and Chen, 2021), utterance intent (Wu et al., 2021), or redundant utterance identification (Feng et al., 2021b). In this work, we propose a dialogue-level annotation describing "*What happened in the dialogue?*" in natural language. The narrated dialogues effectively close the domain gap between pretrain and finetune, as well as between dialogues and summaries. It also helps with data sparsity since the model could effectively transfer knowledge from out-of-domain annotations such as news summarization datasets. The narrative description of dialogues serves as an effective annotator to label coreference, dialogue acts, events, and user intents. Since the narration are in natural language just as the dialogues themselves, we simply replace

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3565–3575
December 7-11, 2022 ©2022 Association for Computational Linguistics



Rephrase

1. Nickola: Have you found it?
2. Sophie: No! Still looking :(
3. Nickola: Check pockets and handbags.
4. Sophie: Checked them all twice already...

Convert the dialog to narrative and formal style for each turn:

1. Nickola asked if Sophie had found it.
2. Sophie said that she had not found it yet and was still looking.
3. Nickola told Sophie to check her pockets and handbags.
4. Sophie said that she had already checked them all twice.

Plot

<Same dialogue transcript as above>

Convert the dialogue into a plot:

This is a conversation between two people who are looking for something. Nickola suggests that Sophie check her pockets and handbags again, and Sophie says that she has already checked them twice.

Event

<Same dialogue transcript as above>

What are the events in this conversation?

-Nickola asks if Sophie has found the item
-Sophie says she has not found the item yet
-Nickola suggests checking pockets and handbags
-Sophie says she has already checked them all twice

Figure 1: Example of the prompt for dialog-to-narrative conversion with InstructGPT model. The narrative types are rephrase, plot and event from the top to bottom.

or concatenate the narration with the dialogue text. As as result, our frame can be plug-and-play with most LLMs without any modification on the model architecture.

We empirically verified the effectiveness of different narrating methods on three benchmark datasets in both supervised and zero-shot settings. Our best narrating method, despite being very simple, outperforming complex and strong baselines on both ROUGE (Lin, 2004) and FactCC (Kryscinski et al., 2020) scores.

2 Methods

We define a dialogue D to be a sequence of turns, where the i-th turn consists of a speaker s_i and a utterance u_i:

$$D = \{(s_1, u_1), \dots, (s_n, u_n)\}$$

A turn in a dialogue is usually represented in the textual format as follows: "$s_i : u_i$", where a colon is used to separate the speaker from the utterance. The format has been treated as the default for some widely used datasets. To better format the turns in a dialogue so that they will be closer to a narrative sentence, we convert the style of a turn using two categories of methods: rule-based transfer and model-based transfer. After conversion, we follow the standard finetune strategy of large-scale language models on the converted dialogue and the original summarization.

2.1 Rule-Based Transfer

Quote To make the dialogue input similar to the web corpus on which large-scaled language models are pretrained, we firstly make dialogue more like written style by simply adding quotation marks '"' to quote the utterance. Specifically, we use the template of <s_i *said,* "u_i".> as the input format for each turn. In the case where the utterance u_i contains a question mark, we change the template to <s_i *asked,* "u_i".> to reflect the fact that the speaker is asking a question.

Pronoun Resolution While quotation marks convert the dialogue default format into written text, the utterance u_i remains unchanged and often contains first and second personal pronouns, e.g. the speakers use "I" to refer to themselves and "you" to refer to other people in the dialogue. To make the dialogue more like a written style, we replace the first and second-person pronouns with the resolved speaker names. Similarly, we use the template of <s_i *said/asked that pron_resol*(u_i)> as the input format for each turn, where *pron_resol*(u_i) is a pronoun resolution function that replaces all the pronouns of a first and second person with resolved speaker names. The replacement will result the lack of verb agreement. Empirically, we found the ungrammatical transcripts after rule-based conversion have little impact on the grammar correctness of the generated summary. We hypothesize that it was because that our summarization models are pretrained and finetuned to always generate grammatically correct English text. Therefore, they can generalize to generate the same even if the input narratives have some grammar errors. The detailed template is shown in appendix A.1.1.

2.2 Model-Based Transfer

Apart from personal pronouns resolution, there are still many characteristics making dialogue summarization challenging. For instance, dialogue transcription contains informal spoken language, which

	ROUGE-1	ROUGE-2	ROUGE-L	FactCC
Pointer Generator	37.27	14.42	34.36	-
Fast Abs RL	41.03	16.93	34.36	-
Transformer	42.37	18.44	39.27	-
DynamicConv	41.07	17.11	37.27	-
DynamicConv + GPT-2 emb	45.41	20.65	41.45	-
DialoGPT	39.77	16.58	38.42	-
DialoGPT-Annotator	53.70	28.79	**50.81**	63.92
MV-BART	53.42	27.98	49.97	-
CODS	52.65	27.84	50.79	-
BART	53.31	28.71	50.61	62.21
w/ quote	53.30	28.38	50.24	62.88
w/ pronoun resolution	52.85	28.12	50.18	63.80
w/ rephrase	51.94	27.11	49.09	62.15
w/ plot	52.60	28.40	50.05	63.06
w/ event	**53.80**	**28.96**	50.76	**64.22**

Table 1: Supervised finetuning result on SAMSum test split. The whole train split is used to finetune a Bart-large model. Baselines performance are taken from works of Wu et al. (2021) and Feng et al. (2021b)

could be informal and noisy. In addition, dialogue is often not well structured with rapid topic changes and unexpected interruptions. To address these challenges, we introduce a model-based method to convert dialogue into well-written narratives that are easier to read and understand.

To the best of our knowledge, there is no existing parallel corpus between dialogue and its narrative equivalence. Therefore, we generate narratives by leveraging the strong zero-shot capability of InstructGPT (Ouyang et al., 2022) to follow instructions for a certain task.

An example of the prompt and model-generated text from InstructGPT is shown in figure 1. To construct the prompt, we first index the turns in the original dialogue with a prefix number starting from 1. The indices are empirically found to help generate comprehensive narratives in our preliminary experiments.

Another trick we applied during prompt construction is the bootstrap text. As shown in the first example of figure 1, the bootstrap text "1. Nickola asked" has two purposes: First, the index constrains the language model to follow the same pattern and generate narratives one by one according to the indices of the original dialogue. Second, "Nickola asked" constrains the following generation to be a third-person narrative. Similar to the rule-based conversion, we replace "asked" with "said" if there is no question mark in the first turn.

To capture various aspects of the dialogue, we propose three model-based narratives: 1) rephrase 2) plot and 3) event. The corresponding instruction and bootstrap text for each type are shown in figure 1. For rephrasing, we aim to narrate the dialog turn-by-turn. For plot narrative, we re-purpose the terminology "plot" in film and play to generate the sequence of interconnected events within the dialogue. For the event narrative, we directly generated the itemized salient events that happened in the dialogue. For rephrasing, we replace the dialogue with the generated text. For the plot and event, since they are shorter and more concise, we concatenate them with the default dialogue format.

3 Experiment

We conduct through experiment on three dialogue summarization datasets: SAMSum (Gliwa et al., 2019), DialogSum (Chen et al., 2021), and ADSC (Misra et al., 2015) under both supervised and zero-shot settings. The properties and statistics of the datasets used in our experiment is shown in table 2.

3.1 Implementation

For supervised finetuning, we used BART-large model (Lewis et al., 2020) from the implementation of HuggingFace [1]. We also tested Pegasus (Zhang et al., 2020b) and T5 (Raffel et al., 2020) in our preliminary experiments, and empirically found that BART achieves the best performance among them. The choice of pretrained language model is also consistent with previous works of Feng et al. (2021b); Chen and Yang (2020); Wu et al. (2021).

For zero-shot experiments, we used the BART-large-CNN as the off-the-shelf summarizer. [2]. The model is initialized on the BART-large model and finetuned on CNN/DailyMail Dataset (See et al., 2017), a widely used news summarization dataset containing over 300k unique news articles. More details of implementation are included in the appendix A.1.

3.2 Evaluation and Baselines

We use standard ROUGE (Lin, 2004) metric as automatic metrics, including ROUGE-1, ROUGE-2, and ROUGE-L. For implementation, we followed Gliwa et al. (2019) to use the py-rouge package [3] with stemming. In addition to ROUGE scores, we introduce FactCC (Kryscinski et al., 2020) as an additional metric for factual correctness. A higher FactCC score means that the system summary is

---

[1] https://huggingface.co/

[2] https://huggingface.co/facebook/bart-large-cnn

[3] https://pypi.org/project/py-rouge/

Dataset	Domain	Train/Dev/Test	# Speakers	# Utterance	# Token (Dialog)	# Token (Summary)	% Novel Words (Summary)
SAMSum	written online dialogue	14732/818/819	2.40	11.20	104.30	24.20	44.0%
DialogSum	spoken daily dialogue	12460/500/500	2.00	9.50	156.00	26.80	34.1%
ADSC	written online debate	0/0/45	2.00	7.50	639.40	157.60	34.9%

Table 2: Properties and statistics of dialog summarization datasets used in the experiments.

	ROUGE-1	ROUGE-2	ROUGE-L	FactCC
Transformer	35.91	8.74	33.50	-
UniLMv2	47.04	21.13	45.04	-
BART	47.51	20.63	45.04	69.83
w/ quote	47.71	21.33	45.25	70.93
w/ pronoun resolution	**47.97**	21.86	**45.61**	**71.83**
w/ rephrase	46.25	20.41	44.31	63.73
w/ plot	47.92	**21.97**	45.29	70.80
w/ event	47.52	20.82	45.10	69.30

Table 3: Supervised finetuning result on DialogSum test split. The whole train split is used to finetune a Bart-large model. Baseline performance are taken from work of Chen et al. (2021)

	Method	ROUGE-1	ROUGE-2	ROUGE-L	FactCC
SAMSum	BART	33.57	10.57	32.41	30.71
	w/ quote	31.75	9.56	30.61	31.32
	w/ pronoun resolution	31.78	10.44	31.19	39.07
	w/ rephrase	32.24	9.51	31.77	50.12
	w/ plot	36.48	13.39	34.19	53.24
	w/ event	**40.07**	**15.24**	**38.56**	**57.14**
DialogSum	BART	27.98	7.2	27.27	31.97
	w/ quote	27.81	7.28	27.09	35.63
	w/ pronoun resolution	29.63	8.54	29.16	52.63
	w/ rephrase	28.66	7.41	28.27	48.50
	w/ plot	31.65	9.62	30.13	45.3
	w/ event	**33.82**	**10.82**	**32.92**	**52.93**
ADSC	BART	27.88	8.46	25.73	16.44
	w/ quote	27.42	7.68	24.61	19.78
	w/ pronoun resolution	**29.52**	7.88	**25.83**	**26.89**
	w/ rephrase	28.96	7.9	25.41	20.89
	w/ plot	28.99	**8.71**	25.01	24.22
	w/ event	29.05	8.63	25.65	**26.89**

Table 4: Zero-shot result on three datasets. We used a BART-large model finetuned on news summarization dataset CNN/DailyMail

more factually consistent with the reference summary.

We refer readers to Chen et al. (2021) for baselines in DialogSum dataset and Wu et al. (2021) for baselines in SAMSum dataset, with an exception of DialoGPT-Annotator Feng et al. (2021b). Similar to our model-based transfer methodology, it uses an external language model, i.e., DigloGPT (Zhang et al., 2020c), to annotate useful information for dialogue summarization, and use BART to finetune on the annotated dialogues.

3.3 Analysis on Results

For supervised results on the SAMSum dataset, using event narratives achieved better performance on all metrics compared with the BART model finetuned on vanilla dialogue inputs. The most salient improvement is on the FactCC score, indicating that adding narratives helps more with factual correctness. Compared with the state-of-art models, such as DialoGPT-Annotator, MV-BART, and CODS, our best-performing model reached comparable or even better performance on both ROUGE and FactCC scores. This empirical result proves the effectiveness of event narratives in the supervised setting. On DialogSum, pronoun resolution narratives performed better than other narrative methods and the BART baseline. One potential reason is that DialogSum only contains two-speaker conversations, therefore pronouns resolved by our rules are 100% correct.

In the zero-shot evaluation, we can see that all narrative variants achieved significant improvement in FactCC scores. Overall, we conclude that the most robust method for zero-shot setting is event narratives. In other words, a news summarizer can benefit most when accessing the itemized events for dialogue summarization. We also notice that the benefit of including narratives is more salient in the zero-shot setting, where there is no in-domain annotated data to help the model to close the domain gap. We further showcase model outputs of our model in appendix A.3.

4 Related Work

To alleviate the domain mismatch and label scarcity, pretraining on dialogue or news domains (Qi et al., 2021; Zou et al., 2021; Khalifa et al., 2021; Zhu et al., 2020), multi-tasking (Liu et al., 2021a; Khalifa et al., 2021), and data argumentation (Chen and Yang, 2021) have shown to be effective for summarization. Other works focus on the structured information of dialogues and model the auxiliary input via graph attention network (Veličković et al., 2018) or the manipulation of transformer attention (Vaswani et al., 2017). The internal structured information includes speaker-utterance relationship (Lei et al., 2021), semantic slot (Zhao et al., 2021), topic (Zhao et al., 2020; Chen and Yang, 2020), and coreference (Liu et al., 2021b). And the external structured information includes commonsense graph from knowledge bases (Xiachong et al., 2021).

Another line of work addresses the challenges of dialogue summarization by directly injecting knowledge into model input. For instance, Liu and Chen (2021) uses personal named entities to control the occurrence of speakers in the summary generation. Feng et al. (2021b) make use of DialogGPT (Zhang et al., 2020c) to annotate keywords, redundant utterance, and topic changes in the dialogue transcript. Wu et al. (2021) use external tools to annotate speaker intent and key phrases as a "sketch" of the dialogue. Our work falls in this category but differs from previous approaches as we attempt to incorporate external knowledge by directly narrating the dialogues in natural language.

5 Conclusion

We propose a general framework to narrate a dialogue into a third-person description for dialogue summarization. We empirically compare different ways of narration and found that the proposed framework improves the performance, especially the factual correctness of the generated summary, for both supervised and zero-shot settings in three benchmark datasets. The improvement is most consistent when including the salient events in the dialogue as narrative. The resulting summarization model surpasses existing strong baselines on SAMSum and DialogSum datasets.

6 Limitations

We have not explored towards a principled way to combine different narratives to achieve better performance, as well as the combination of narrating dialogues with other dialogue summarization techniques. We leave these directions to future work.

References

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021a. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021b. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics.

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.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Yuejie Lei, Fujia Zheng, Yuanmeng Yan, Keqing He, and Weiran Xu. 2021. A finer-grain universal dialogue semantic structures based model for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1354–1364, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021a. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021b. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G. Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. GupShup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gu°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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

MengNan Qi, Hao Liu, YuZhuo Fu, and Ting Liu. 2021. Improving abstractive dialogue summarization with hierarchical pretraining and topic segment. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1121–1130, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. Association for Computational Linguistics.

Feng Xiachong, Feng Xiaocheng, and Qin Bing. 2021. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 964–975, Huhhot, China. Chinese Information Processing Society of China.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020b. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lulu Zhao, Weihao Zeng, Weiran Xu, and Jun Guo. 2021. Give the truth: Incorporate semantic slot into abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2435–2446, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021a. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021b. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1462–1471.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

The total number of steps was set to 8000 with linear learning scheduler peaked at step 800. The learning rate is set to be $3e-5$ and the batch size is set to be 8, with a gradient accumulation of 4 steps. We tuned a weight decay coefficient at from $[0.1, 0.01, 0.001]$ and a label smoothing factor at from $[0.0, 0.1]$ using grid search and selected weight decay to be 0.001 and label smoothing factor to be 0.1 on the validation set performance of SAMSum dataset with its original dialogue input. The beam size is set to 4 for decoding. For both finetuning and inference, we used the pipeline from Hugging Face (Wolf et al., 2020). [4] All experiments are done on a cluster with 8 NVIDIA V100 32GB GPUs.

A.1.1 Pronoun Resolution

For pronoun identification, we used the part-of-speech pipeline from spaCy package [5]. The rule of replacing pronouns with speaker names is specified in table 5.

Pronouns Word	Replacement
I, me	<speaker>
you, u	<prev_speaker>
we, us	<speaker>and <prev_speaker>
my	<speaker>'s
your, ur	<prev_speaker>'s
our	<speaker>and <prev_speaker>'s

Table 5: Resolution mapping of first and second pronouns to speaker names. Placeholder <speaker>represents the speaker's name of the current turn. <prev_speaker>represents the speaker's name of the previous turn. If the current turn is the first turn, then <prev_speaker>is the current speaker.

A.1.2 Rephrase Generation

For rephrase generation, we first convert dialogue to narrative equivalence with InstructGPT. Secondly, we filter the generated narratives based on several criteria to make sure that resulting parallel corpus has a high quality. Specifically, we compute the length of the narratives in word and divide it by the length of the dialogue. If the ratio is smaller than a threshold, the narrative tends to lose some key information of the dialogue and is therefore discarded. In our experiment, we manually inspect some of the generated narratives and set this threshold to be 1.2. Similarly, we also set a upper bound for this ratio to be 5, preventing the narrative to be too verbose when compared with the original dialog. In filtering process, we found the upper bound is very rarely reached. At last, we use the filtered parallel corpus from SAMSum dataset to finetune a BART-Large model and use it to convert dialogue in the all downstream datasets.

A.2 FactCC Modification

The original FactCC takes the source text and the claim as inputs to verify whether the claim is supported by the source text. For dialogue summarization, the input dialogue text has a domain gap between the training data of FactCC model, therefore, we input the reference summary instead in order to determine if the system predicted summary is well supported by the reference summary. Similarly, we also input the system and reference summary in the reverse order and averaged the two scores for each testing example.

A.3 Example Model Outputs

We show the example outputs from baseline BART model and narrative model with events in figures 2, 3, and 4 for SAMSum, DialogSum and ADSC dataset respectively. We can see that the event narratives helps BART model to generate more detailed summaries than using the dialogue input alone. The quality improvement is significant in the zero-shot setting, where BART w/ event could still generate high-quality summaries, the vanilla zero-shot model tends to copy turns from the dialogue or produce hallucinations.

---

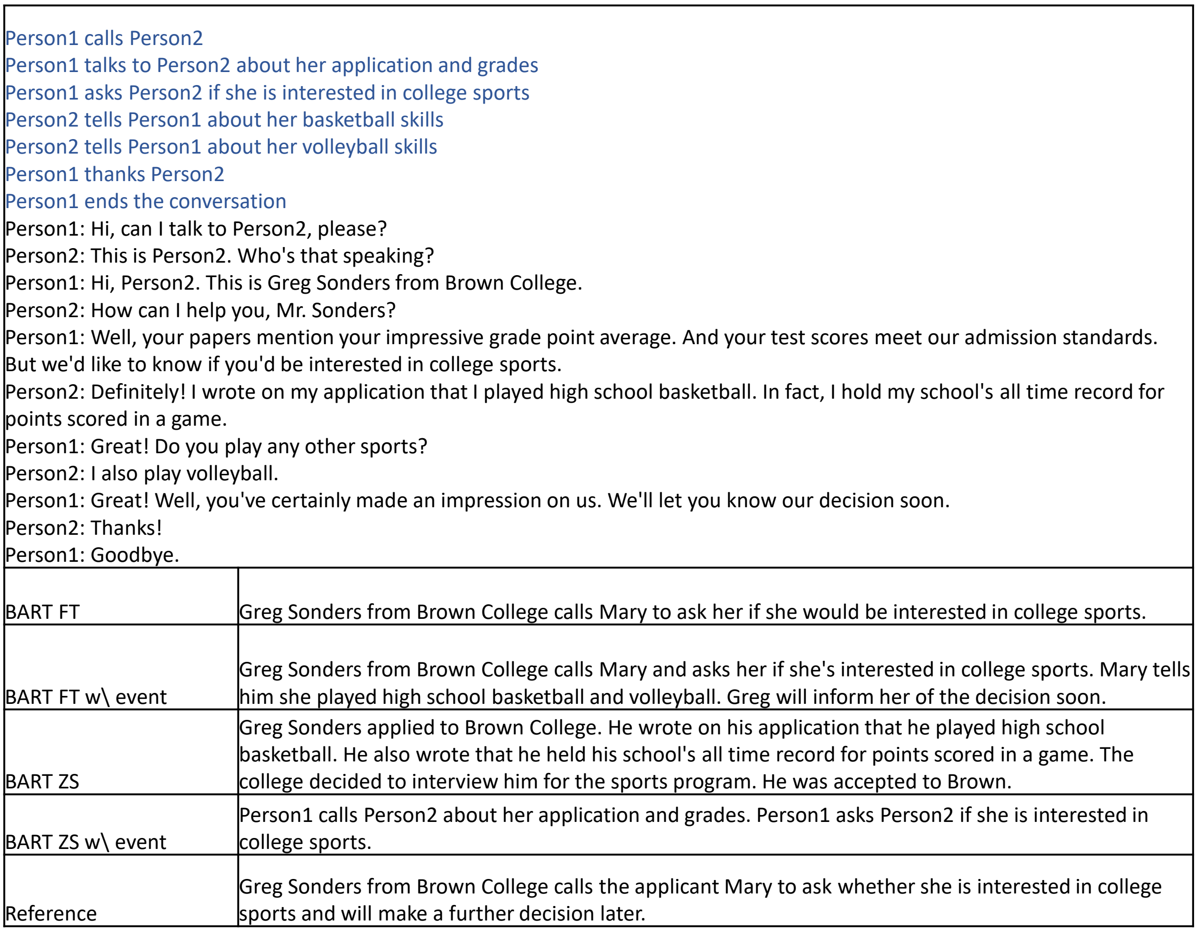
[4] Source code from https://github.com/huggingface/transformers/tree/v4.9.1/examples/pytorch/summarization

[5] https://spacy.io/

Wanda asking Gina to help her plan a party
Gina agreeing to help
Wanda making a list of things she needs for the party
Gina agreeing to go grocery shopping with Wanda
Wanda: Let's make a party!
Gina: Why?
Wanda: beacuse. I want some fun!
Gina: ok, what do u need?
Wanda: 1st I need too make a list
Gina: noted and then?
Wanda: well, could u take yours father car and go do groceries with me?
Gina: don't know if he'll agree
Wanda: I know, but u can ask :)
Gina: I'll try but theres no promisess
Wanda: I know, u r the best!
Gina: When u wanna go
Wanda: Friday?
Gina: ok, I'll ask

BART FT	Wanda wants to make a party. She wants Gina to go shopping with her on Friday.
BART FT w\ event	Wanda wants to make a party. Gina will help her with that. She will go shopping with Wanda on Friday.
BART ZS	Wanda: Let's make a party! "I want some fun!" "I'll try but theres no promisess"
BART ZS w\ event	Wanda asks Gina to help her plan a party. Gina agrees to help. Wanda makes a list of things she needs for the party.
Reference	Wanda wants to throw a party. She asks Gina to borrow her father's car and go do groceries together. They set the date for Friday.

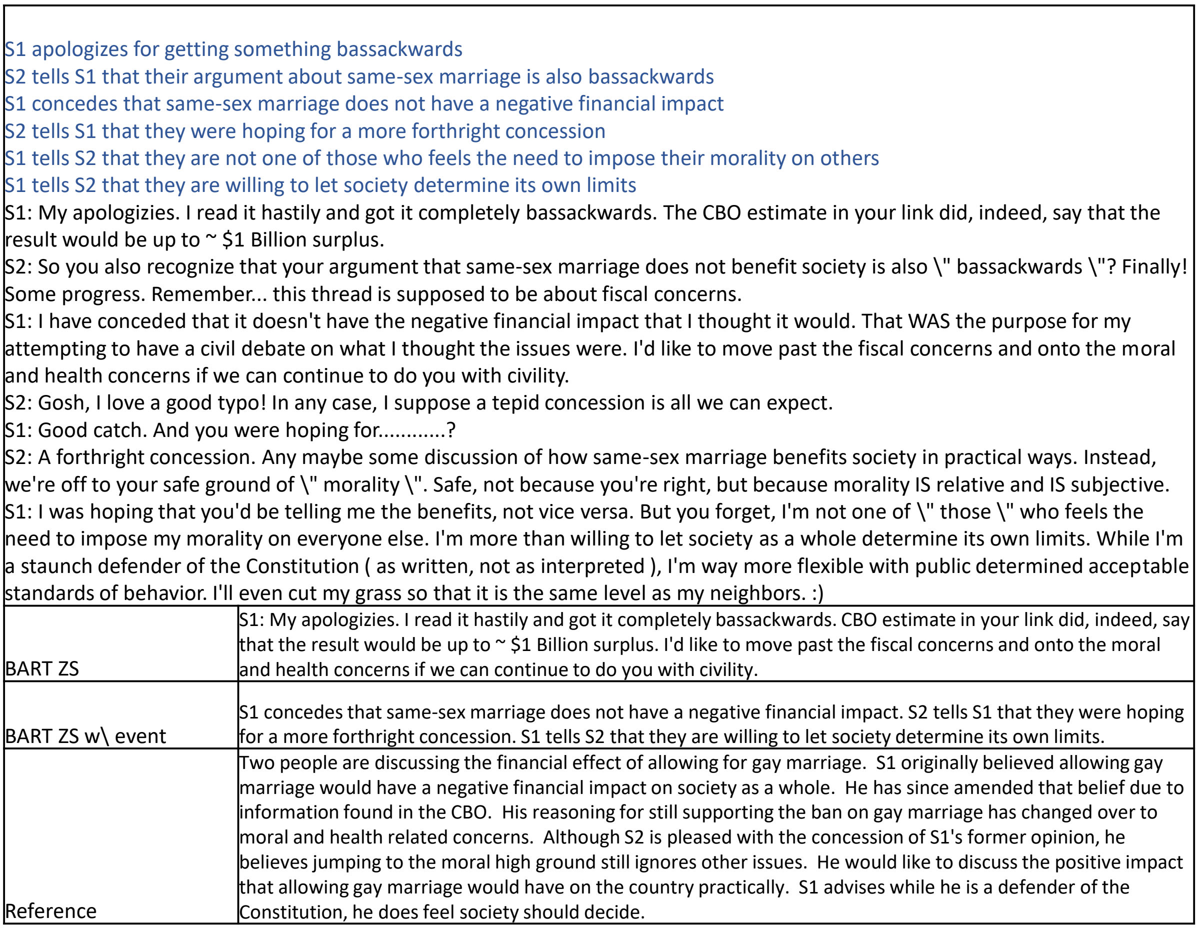
Figure 2: Example input and output on SAMSum dataset of BART with event narratives (in blue color). "FT" means finetuned model outputs, and "ZS" means zero-shot model outputs.



Person1 calls Person2
Person1 talks to Person2 about her application and grades
Person1 asks Person2 if she is interested in college sports
Person2 tells Person1 about her basketball skills
Person2 tells Person1 about her volleyball skills
Person1 thanks Person2
Person1 ends the conversation
Person1: Hi, can I talk to Person2, please?
Person2: This is Person2. Who's that speaking?
Person1: Hi, Person2. This is Greg Sonders from Brown College.
Person2: How can I help you, Mr. Sonders?
Person1: Well, your papers mention your impressive grade point average. And your test scores meet our admission standards. But we'd like to know if you'd be interested in college sports.
Person2: Definitely! I wrote on my application that I played high school basketball. In fact, I hold my school's all time record for points scored in a game.
Person1: Great! Do you play any other sports?
Person2: I also play volleyball.
Person1: Great! Well, you've certainly made an impression on us. We'll let you know our decision soon.
Person2: Thanks!
Person1: Goodbye.

BART FT	Greg Sonders from Brown College calls Mary to ask her if she would be interested in college sports.
BART FT w\ event	Greg Sonders from Brown College calls Mary and asks her if she's interested in college sports. Mary tells him she played high school basketball and volleyball. Greg will inform her of the decision soon.
BART ZS	Greg Sonders applied to Brown College. He wrote on his application that he played high school basketball. He also wrote that he held his school's all time record for points scored in a game. The college decided to interview him for the sports program. He was accepted to Brown.
BART ZS w\ event	Person1 calls Person2 about her application and grades. Person1 asks Person2 if she is interested in college sports.
Reference	Greg Sonders from Brown College calls the applicant Mary to ask whether she is interested in college sports and will make a further decision later.

Figure 3: Example input and output on DialogSum dataset of BART with event narratives (in blue color). "FT" means finetuned model outputs, and "ZS" means zero-shot model outputs.



S1 apologizes for getting something bassackwards
S2 tells S1 that their argument about same-sex marriage is also bassackwards
S1 concedes that same-sex marriage does not have a negative financial impact
S2 tells S1 that they were hoping for a more forthright concession
S1 tells S2 that they are not one of those who feels the need to impose their morality on others
S1 tells S2 that they are willing to let society determine its own limits

S1: My apologizies. I read it hastily and got it completely bassackwards. The CBO estimate in your link did, indeed, say that the result would be up to ~ $1 Billion surplus.

S2: So you also recognize that your argument that same-sex marriage does not benefit society is also \" bassackwards \"? Finally! Some progress. Remember... this thread is supposed to be about fiscal concerns.

S1: I have conceded that it doesn't have the negative financial impact that I thought it would. That WAS the purpose for my attempting to have a civil debate on what I thought the issues were. I'd like to move past the fiscal concerns and onto the moral and health concerns if we can continue to do you with civility.

S2: Gosh, I love a good typo! In any case, I suppose a tepid concession is all we can expect.

S1: Good catch. And you were hoping for...........?

S2: A forthright concession. Any maybe some discussion of how same-sex marriage benefits society in practical ways. Instead, we're off to your safe ground of \" morality \". Safe, not because you're right, but because morality IS relative and IS subjective.

S1: I was hoping that you'd be telling me the benefits, not vice versa. But you forget, I'm not one of \" those \" who feels the need to impose my morality on everyone else. I'm more than willing to let society as a whole determine its own limits. While I'm a staunch defender of the Constitution (as written, not as interpreted), I'm way more flexible with public determined acceptable standards of behavior. I'll even cut my grass so that it is the same level as my neighbors. :)

BART ZS	S1: My apologizies. I read it hastily and got it completely bassackwards. CBO estimate in your link did, indeed, say that the result would be up to ~ $1 Billion surplus. I'd like to move past the fiscal concerns and onto the moral and health concerns if we can continue to do you with civility.
BART ZS w\ event	S1 concedes that same-sex marriage does not have a negative financial impact. S2 tells S1 that they were hoping for a more forthright concession. S1 tells S2 that they are willing to let society determine its own limits.
Reference	Two people are discussing the financial effect of allowing for gay marriage. S1 originally believed allowing gay marriage would have a negative financial impact on society as a whole. He has since amended that belief due to information found in the CBO. His reasoning for still supporting the ban on gay marriage has changed over to moral and health related concerns. Although S2 is pleased with the concession of S1's former opinion, he believes jumping to the moral high ground still ignores other issues. He would like to discuss the positive impact that allowing gay marriage would have on the country practically. S1 advises while he is a defender of the Constitution, he does feel society should decide.

Figure 4: Example input and output on ADSC dataset of BART with event narratives (in blue color). "ZS" means zero-shot model outputs.