Post-Training Dialogue Summarization using Pseudo-Paraphrasing

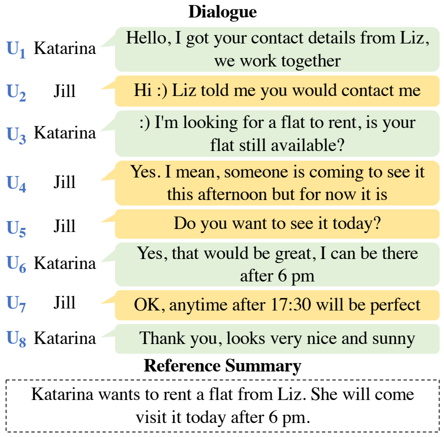
Qi Jia[1], **Yizhu Liu**[1], **Haifeng Tang**[2], **Kenny Q. Zhu**[3*]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]China Merchants Bank Credit Card Center, Shanghai, China
[1]{Jia_qi, liuyizhu}@sjtu.edu.cn
[2]thfeng@cmbchina.com
[3]kzhu@cs.sjtu.edu.cn

Abstract

Previous dialogue summarization techniques adapt large language models pretrained on the narrative text by injecting dialogue-specific features into the models. These features either require additional knowledge to recognize or make the resulting models harder to tune. To bridge the format gap between dialogues and narrative summaries in dialogue summarization tasks, we propose to post-train pretrained language models (PLMs) to rephrase from dialogue to narratives. After that, the model is fine-tuned for dialogue summarization as usual. Comprehensive experiments show that our approach significantly improves vanilla PLMs on dialogue summarization and outperforms other SOTA models by the summary quality and implementation costs.[1]

1 Introduction

Dialogue summarization is a specialized summarization task that takes a series of utterances from multiple speakers in the first person as input, and outputs fluent and concise summaries in third persons as shown in Figure 1. Different from previous monologue inputs such as news (Narayan et al., 2018) and scientific publications (Cohan et al., 2018), dialogues are always less well-organized. They usually contain complicated reference relations, inconsecutive inter-utterance dependencies, informal expressions, and so on, making dialogue summarization a more challenging task.



Dialogue

U_1 Katarina: Hello, I got your contact details from Liz, we work together

U_2 Jill: Hi :) Liz told me you would contact me

U_3 Katarina: :) I'm looking for a flat to rent, is your flat still available?

U_4 Jill: Yes. I mean, someone is coming to see it this afternoon but for now it is

U_5 Jill: Do you want to see it today?

U_6 Katarina: Yes, that would be great, I can be there after 6 pm

U_7 Jill: OK, anytime after 17:30 will be perfect

U_8 Katarina: Thank you, looks very nice and sunny

Reference Summary

Katarina wants to rent a flat from Liz. She will come visit it today after 6 pm.

Figure 1: An example from SAMSum dataset.

The most obvious characteristic of this task is the difference in the format and language styles between dialogue and its narrative summary. Liu, Shi and Chen (2021b) mentioned that coreference resolution models trained on general narrative text underperforms by about 10% on dialogue corpus, demonstrating the inherent gap between dialogue and narrative text. As a result, popular PLMs such as BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020a) which excel on news summarization perform mediocrely on dialogue summarization.

To narrow this gap, previous work on dialogue summarization mainly resort to injecting dialogue features into PLMs to enhance dialogue understanding. These features include dialogue acts (Goo and Chen, 2018), topic transitions (Chen and Yang, 2020), coreference relations (Liu et al., 2021b), discourse graphs (Chen and Yang, 2021), etc, leading to the rule-based conversion from dialogues to plain text (Ganesh and Dingliwal, 2019). However, they suffer from three weaknesses. First, collecting or extracting these features becomes an additional step in the summarization pipeline, complicating the inference procedure at runtime. Second, oracle feature labels are hard to collect and errors can propagate from wrong labels to poor summaries. Third, additional layers or more encoders are required to incorporate features into PLMs, increasing the GPU memory footprint both during training and inference.

A more natural way to bridge this gap is to give the model more dialogue-narrative pairs to train on.

[*] The corresponding author.

[1] Our code and results are publicly available at https://github.com/JiaQiSJTU/DialSent-PGG.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1660 - 1669
July 10-15, 2022 ©2022 Association for Computational Linguistics

Due to the scarcity of dialogue summarization data, one approach (Zhu et al., 2020) is to convert other text summarization pairs into dialogue to summary pairs via some template, but such work requires additional data [2].

In this paper, we propose an alternative approach that doesn't use any more data than the original dialogue summarization dataset. We convert each existing data pair into many "**pseudo-paraphrase**" pairs between a dialogue and a narrative sentence. Then we post-train a pre-trained seq2seq language model using a **prefix-guided generation** (PGG) task on the augmented paraphrase dataset. After that, the post-trained model is further fine-tuned as usual for dialogue summarization. To this end, no human efforts on crafting complicated rules or hyper-parameter tuning, or additional memory costs, as well as additional training data, is required. In sum, our contributions are:

- We propose a novel and effective post-training process to close the format and linguistic style gap between dialogues and narrative texts (§ 2).
- PGG with pseudo-paraphrase pairs requires no extra training data or labeling tools for features extractions (§ 3.2).
- Extensive experiments show that the proposed approach compares favorably with current SOTA models using less human efforts and computational costs (§ 3.3).

2 Approach

The training of a dialogue summarization model is divided two stages: post-training and fine-tuning. The model can be any seq-to-seq PLMs and it remains unchanged except for the parameters which are updated stage by stage. We will elaborate on the post-training stage in the rest of this section.

2.1 Pseudo-paraphrase Dataset Construction

We construct rephrasing datasets from the dialogue summarization dataset itself. The original dialogue summarization dataset (**DSum**) is made up of dialogue-summary (D-S) pairs. Each dialogue D is a sequence of utterances and can be concatenated into a whole sequence:

$$D = \{U_1, U_2, ..., U_T\} = \{x_1, \ldots, x_n\} \qquad (1)$$

Each turn U_t is in the form of $[r_t\text{: } u_t]$, where r is a speaker and u is the actual utterance.

Our goal is to create more dialogue to narration kind of paraphrasing pairs. The most intuitive approach is to divide S into sentences, and pair each sentence to D. We call such pairs "pseudo-paraphrases" because the output sentence (which we call p) isn't exactly the paraphrase of the whole input, but rather part of the input.

However, doing this poses two challenges: 1) S is a coherent piece of text, and its sentences may depend on each other, so a single sentence p out of it may not stand by itself; 2) one D will be paired with several different p, and it is hard for the model to distinguish the meaning of these pairs.

Datasets	**Input**	**Output**
DSum	$U_{1\sim8}$	Katarina wants to rent a flat from Liz. She will come visit it today after 6 pm.
DialSent	$U_{1\sim8}$	Katarina *wants* to rent a flat from Liz.
	$U_{1\sim8}$	Katarina *will come* visit it today after 6 pm.

Table 1: Example pseudo-paraphrase pairs generated from the example in Figure 1. One pair in DSum becomes two pairs in DialSent. The prefix tokens determined by linguistic features, NOUN and ROOT, are underlined and italic respectively.

To solve 1) we apply coreference resolution[3] on S and convert every personal pronoun in it to the full reference first, before splitting the summary S into sentences. Sentence with fewer than 3 words (e.g., "Ally agree") are discarded since it carries too little information. The set of data pairs thus created is called (**DialSent**). An example is in Table 1.

To tackle 2), one obvious thought is to further split D into sets of sentences in which each set corresponds to a sentence p in the summary. However, our extensive experiments (see Appendix C) showed that none of the straight-forward heuristics work well to establish such alignments. This is mainly due to the fact that dialogue utterances are highly dependent. Thus, splitting operations are not optimal. Instead of changing D, we decide to use the pseudo-paraphrases directly but introduce a prefix-guided generation task to guide the model learning to extract relevant information from D.

2.2 Prefix-guided Generation Task

Summarization for dialogues focuses on analyzing "who-did-what" storylines (Chen and Yang, 2021) and the beginning of each summary sentence are

[2] More related work is in Appendix A.

[3] We use `https://spacy.io/`.

usually different speakers or the same speaker doing different things. As a result, using the prefix made up of "who" or "who-did" can help to select the related information from dialogues or plan the content to be generated.

In other words, we take the inspiration from content planning (Narayan et al., 2021; Wu et al., 2021). When training, the first few tokens of p are provided as prefix to the decoder. This prefix serves as an information selection hint to the model so it is easier to learn why that particular p should be generated. The losses are calculated between the generated tokens and reference tokens after the prefix as shown in Figure 2.

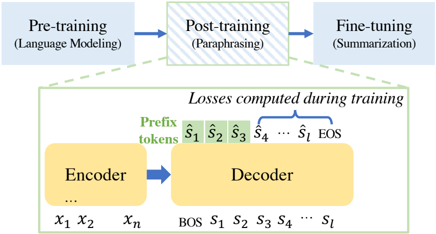

Pre-training
(Language Modeling)
Post-training
(Paraphrasing)
Fine-tuning
(Summarization)
Losses computed during training
Prefix tokens
$\hat{s}_1$ $\hat{s}_2$ $\hat{s}_3$ $\hat{s}_4$ ⋯ $\hat{s}_l$ EOS
Encoder
Decoder
…
x_1 x_2 x_n
BOS s_1 s_2 s_3 s_4 ⋯ s_l

Figure 2: A illustration of our approach. BOS and EOS stand for begin and end of the sequence.

Let $p = \{s_1, \ldots, s_l\}$. Our prefix-guided training task is a vanilla auto-regressive generation task minimizing the negative log-likelihood of p:

$$L = -\frac{1}{l-a}\sum_{t=a}^{l}\log P(s_t|s_{<t}, H^d) \qquad (2)$$

where a is the number of prefix tokens. H^d is the output hidden vectors of the encoder with input D.

There are various ways to determine the prefix length a. We can take a fixed length, a random length or a prefix up to a certain linguistic feature such as NOUN, VERB or ROOT. The exact linguistic feature to use is a dataset-dependent hyperparameter and can be tuned by the validation set. Examples of prefix tokens is marked in Table 1.

3 Evaluation

We first present the experimental setups, then conduct an ablation study to determine the proper prefix in PGG training, before our main results. More implementation details are in Appendix B.

3.1 Experimental Setup

We implement our experiments on **SAMSum** (Gliwa et al., 2019) and **DialSumm** (Chen et al., 2021), whose statistics are listed in Table 2.

Datasets	**Variation**	**Train/Val/Test**	**IW**	**OW**	**CR**
SAMSum	DSum	14,731/818/819	124.10	23.44	0.25
	DialSent	29,757/1,654	149.93	11.93	0.13
DialSumm	DSum	12,460/500/500	187.52	31.02	0.18
	DialSent	22,407/840	214.00	17.78	0.10

Table 2: Statistics of dialogue summarization datasets. IW, OW and CR represent the number of input words, the number of output words and compression ratio (OW/IW) respectively.

We compare our method with these baselines. **Lead-3** and **Longest-3** are simple rule-based baselines that extract the first or the longest 3 utterances in a dialogue as the summary respectively. **PGN** (See et al., 2017), **Fast-Abs** (Chen and Bansal, 2018), and **PEGASUS** (Zhang et al., 2020a) are well-known models for text summarization. **BART** (Lewis et al., 2020) is a general PLM and performs well after fine-tuning. **CODS** (Wu et al., 2021), **Multi-view** (Chen and Yang, 2020) and **DialoBART** (Feng et al., 2021b) are the SOTA models designed for dialogue summarization.

We evaluate both automatically and by human. For **automatic evaluation**, we use Rouge-1, 2, and L (Lin, 2004) F1-scores [4]. Following Feng et al. (2021b), we adopt the same Rouge evaluation tool and compute between reference summaries and generated summaries. For DialSumm, we use maximum rouge scores among references for each sample. For **human evaluation**, we three proficient English speakers to evaluate 100 random samples from SAMSum. Each original dialogue and its reference summary are shown with generated summaries in a random order simultaneously. Showing summaries from different approaches together helps humans do comparisons between them. Following Chen and Yang (2020) and Liu et al. (2021b), each summary is scored on the scale of $[2, 0, -2]$, where 2 means concise and informative, 0 means acceptable with minor errors, and -2 means unacceptable. The final scores are averaged among annotators. We also ask human annotators to label the error types in the summary. We consider the following 4 error types: **Mis**sing important contents, **Red**undant content, **Cor**eference mismatches, and **Rea**soning error. Rea and Cor concentrate on comparisons to the dialogue, and the rest two focus on comparisons to the reference. We determine the error for each case by majority voting, and count the errors of each model.

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

Models	Rouge-1	Rouge-2	Rouge-L
SAMSum			
DSum-VG	51.48	27.27	49.45
DSum-PGG	52.52	27.51	49.03
DialSent-VG	52.16	27.79	49.41
DialSent-PGG	**53.54**	**28.91**	**50.21**
DialSumm			
DSum-VG	53.15	28.86	51.48
DSum-PGG	53.27	28.64	51.69
DialSent-VG	52.99	29.14	51.40
DialSent-PGG	**54.73**	**30.47**	**53.46**

Table 3: Ablations on DialSent with PGG task.

3.2 Ablations Study

We conduct ablations to verify the effectiveness of post-training on DialSent with PGG, including post-training on DSum with PGG task (DSum-PGG), DSum with vanilla generation task (DSum-VG), and DialSent with vanilla generation task (DialSent-VG) in Table 3. The results of DSum-VG drop, indicating that fine-tuning for BART on DSum with early-stop is enough. Post-training with the same data and task leads to overfitting. DialSent-PGG performs best for two reasons. Compared with DialSent-VG, the prefix solves one-to-many mappings between a dialogue and summary sentences, so that the same dialogue can lead to different generations. On the other hand, the prefix can manipulate the selection within a short sentence but is not strong enough to direct content in multiple sentences. Thus, DialSent-PGG learns more cross-format paraphrasing ability and performs better.

Models	Rouge-1	Rouge-2	Rouge-L
SAMSum			
w/o	52.16	27.79	49.41
const	51.71	27.34	49.25
random	52.32	27.99	49.68
Ling-Noun	**53.54**	**28.91**	50.21
DialSumm			
w/o	52.99	29.14	51.40
const	53.29	29.57	52.10
random	53.82	29.88	52.43
Ling-Root	**54.73**	**30.47**	**53.46**

Table 4: Ablations on prefix designs for PGG.

We try several choices of prefix length: (1) **W/O**: without any prefix. (2) **Const**: Constant length set to 2 and 3 for SAMSum and DialSumm respectively, since a person's name is 1.69 ± 0.69 tokens long on average [5]. (3) **Random**: set by uniform sampling from a range of numbers. We set the range to $1 \sim 3$ and $2 \sim 4$ for the two datasets respectively. (4) **Ling**: using the validation set, we determined that Noun and Root are the best choice for the two datasets, respectively. In this way, the number of prefix tokens for SAMSum and DialSum are 1.90 ± 1.10 and 3.55 ± 1.24.

In Table 4, Ling performs the best among these variants. The actual linguistic feature to use may vary from dataset to dataset though. The remaining experiments will be conducted using PGG-Ling.

Models	Rouge-1	Rouge-2	Rouge-L
SAMSum			
Lead-3	31.41	8.68	30.38
Longest-3	32.46	10.27	29.92
PGN	40.08	15.28	36.63
Fast-Abs	41.95	18.06	39.23
PEGASUS	50.50	27.23	49.32
BART†	52.06	27.45	48.89
CODS	52.65	27.84	50.79
Multi-view	53.42	27.98	49.97
DialoBART	**53.70**	28.79	**50.81**
DialSent-PGG†	53.54	**28.91**	50.21
DialSumm			
Lead-3	31.15	10.08	30.68
Longest-3	27.00	9.41	25.31
BART†	53.01	29.18	51.34
DialoBART†	53.26	29.58	52.01
DialSent-PGG†	**54.73**	**30.47**	**53.46**

Table 5: Dialogue summarization results compared with baselines. † represents the models implemented by ourselves. Underlined scores are statistically significantly better than BART with $p < 0.05$ based on t-test.

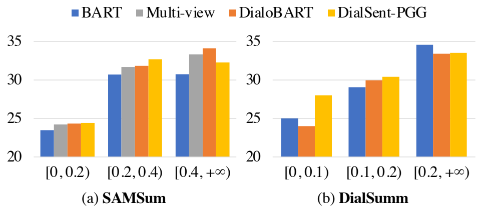
3.3 Comparison to SOTA Models

Automatic Evaluation: Our model DialSent-PGG performs competitively against other models on SAMSum and significantly better than the peers on DialSumm. It improves 1.5 on Rouge scores over BART for both datasets, while DialoBART achieves less gains on DialSumm. Based on Table 1, DialSumm is a more difficult dataset with lower compression ratios. Our model performs better on samples with lower CR, i.e. more compressed samples, as shown in Figure 3, thus differences between DialSent-PGG and DialoBART are more obvious on DialSumm. A simple case study is shown in Table 6. Multi-view faces the repetition problem as it takes the dialogue as input twice with two encoders. DialoBART has reasoning errors because it regards "William" as a keyword. DialSent-PGG instead generates a concise and correct summary. More cases are in Appendix D.

Human Evaluation: The overall human scores on BART, Multi-view, DialoBART and DialSent-PGG are 0.35, 0.40, 0.43 and 0.55 respectively. The Fleiss Kappa among three annotators is 0.39 [6].

[5] DialSumm normalizes speaker names into "#Person1#" resulting in more tokens.

[6] Fleiss Kappa between 0.4 and 0.6 is considered moderate.



(a) **SAMSum**
(b) **DialSumm**

Figure 3: Comparison for models on samples with different CR. X-axis represents the ranges for CR(%). Y-axis is the Rouge-2 F1(%).

Dialogue	William: are you still angry? Emilia: YES William: :(
Multi-view	Emilia is still angry *and still angry*.
DialoBART	*William and* Emilia are still angry.
DialSent-PGG	Emilia is still angry.

Table 6: A case from SAMSum. *Errors* are in italic.

The latter three models all improve BART, with DialSent-PGG topping the ranks.

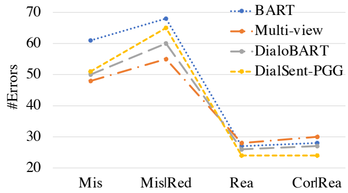


Figure 4: Error analysis on SAMSum.

For error analysis, the Fleiss Kappa for Mis, Red, Cor and Rea are 0.55, 0.10, 0.26, 0.42 respectively. The agreement on Red is lower because identifying unimportant information is hard. The agreement on Cor is fair due to undistinguishable errors. For example, mismatching of a person and an event among multiple utterances can be either a Cor or a Rea. Besides, Red always leads to Mis. So, we divide the error types into two groups and merge them with "OR" logical operation within each group. The Fleiss Kappa for Mis|Red and Cor|Rea are 0.45 and 0.46. We show error types with the agreement larger than 0.40 in Figure 4.

Multi-view performs better on content selection and DialSent-PGG performs better on reasoning and coreference understanding, while DialoBART lies in between. Fewer errors on Rea and Cor|Rea reflect that our approach successfully narrows the understanding gap. Because references are not the only good summary, high missing content doesn't mean that the generated summary is unacceptable. As a result, the model with fewer Cor|Rea errors receives higher overall score.

Implementation Costs: We compare the implementation costs between our approach and two state-of-the-art models, i.e. Multi-view and DialoBART, in Table 7. Although explicitly injecting features for dialogue understanding is effective, labels for these features are hard to collect and implementation costs for these approaches on a new dataset are high. Multi-view and DialoBART proposed doing labeling automatically with unsupervised algorithms or language models. However, these labeling approaches bring extra hyper-parameters which are different between datasets and need to be found by trial and error. If we use the same keywords extraction ratio, similarity threshold and topic segmentation ratio from SAMSum directly, the results on DialSumm are only 50.61/26.67/49.06 (Rouge-1/2/L). We searched for the best combination of hyper-parameters following their paper and did 14 trials, while applying our approach on DialSumm only need 4 trials.

On the other hand, injecting features increases the requirement of GPU memory. With the same training parameters(max tokens=1024, batch size=1, gradient checkpointing=False), Multi-view with double-encoder design encounters an out-of-memory error on RTX 2080Ti with 11G GPU memory. DialoBART occupies around 10.36G since it lengthens the dialogue with additional annotations. DialSent-PGG only occupies 9.87G during post-training for recording the length of the prefix, and 9.65G during fine-tuning which is the same as vanilla BART. In a word, our approach costs less for implementation.

Models	**Mem**	**#HP**	**#Tri**	**#St**
Multi-view	OOM	5	-	-
DialoBART	10.36G	3	14	38.61k
DialSent-PGG	9.87G/9.65G	1	4	19.32k

Table 7: The upper-bound of GPU memory footprint (Mem), newly introduced hyper-parameter counts (#HP), the number of trails (#Tri) and total training steps (#St) for implementing different models.

4 Conclusion

We propose to post-train dialogue summarization models to enhance their cross-format rephrase ability by prefix-guided generation training on dialogue-sentence pseudo-paraphrases, and get promising results. Creating self-supervised tasks for cross-format post-training and incorporating compatible features for downstream fine-tuning are plausible future directions.

Acknowledgement

This research is partially supported by NSFC Grant No. 91646205, and SJTU-CMBCC Joint Research Scheme.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

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.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391.

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.

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.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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.

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.

Prakhar Ganesh and Saket Dingliwal. 2019. Restructuring conversations using discourse relations for zero-shot abstractive dialogue summarization. *arXiv preprint arXiv:1902.01615*.

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.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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, 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.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021. Planning with entity chains for abstractive summarization. *arXiv preprint arXiv:2104.07606*.

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*, 21:1–67.

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.

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. 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. 2020b. 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.

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.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. 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. *arXiv preprint arXiv:2004.02016*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Related Work

Dialogue summarization and pretrained language models are discussed as follows.

Dialogue Summarization: A growing number of works have been proposed for dialogue summarization in recent years. In this work, we mainly refer to the chat summarization defined in (Feng et al., 2021a). Previous works widely explore dialogue features explicitly and input them as known labels to enhance the dialogue understanding ability of summarization models. Features, including dialogue acts (Goo and Chen, 2018), topic transitions (Chen and Yang, 2020), discourse dependencies (Chen and Yang, 2021), coreference relations (Liu et al., 2021b), argument graphs (Fabbri et al., 2021), semantic structures or slots (Lei et al., 2021; Zhao et al., 2021), etc. are carefully designed and collected by transferring tools pre-trained on other corpus or unsupervised methods with multiple hyper-parameters. These work also modify the basic transformer-based models with additional encoders (Chen and Yang, 2020) or attention layers (Chen and Yang, 2021; Liu et al., 2021b; Lei et al., 2021; Zhao et al., 2021) to utilize the injected features. Liu et al. (2021a) propose a contrastive learning approach for dialogue summarization with multiple training objectives. They also introduce a number of hyper-parameters for contrastive dataset construction and balancing among those objectives.

Pretrained Language Models: Previous pretrained seq-to-seq models can be divided into two categories by training data formats. One is models pretrained on narrative text, such as BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020a), and T5 (Raffel et al., 2020). They use training data from Wikipedia, BookCorpus (Zhu et al., 2015) and C4 (Raffel et al., 2020). These models show great potentials for tasks such as translation and story ending generation. The other is models pretrained on dialogue, such as DialoGPT (Zhang et al., 2020b) and PLATO (Bao et al., 2020). Their training data are general-domain dialogues, such as Reddit (Henderson et al., 2019) and Twitter (Cho et al., 2014). These models work for dialogue response selection and generation tasks. All of the above models are trained to exploit language features within the same data format, with pre-training tasks such as masked token/sentence prediction and utterance permutation. Pretraining with cross-format data hasn't been researched so far. As a first step, we focus on narrowing the gap by learning to rephrase unidirectionally from dialogue to narratives.

B Implementation Details

We use BART[7] as our basic language model. For both post-training and fine-tuning, the speakers and utterances of each dialogue are concatenated into a single sequence and truncated to the first 1024 tokens. The learning rate is set to $3e{-}5$ with weight decay equaling 0.01. The number of warmup steps is 500 and dropout is 0.1. The model is tested on the corresponding validation set after each training epoch and the early-stop is activated if there is no improvement in the Rouge-2 F1 score. The early-stop and maximum training epochs are set to 3 and 10. During inference, i.e., validation and testing, the beam size is set to 4 with length penalty equaling 1.0 and no-repeat-n-gram size equaling 3. The minimum and maximum lengths are set to the corresponding lengths of the reference summaries based on statistics of each dataset, allowing for free-length text generation. Besides, for the inference on the validation set during the post-training stage, we also set the first 3 tokens as the known prefix. This constant number enables a fair comparison of performances on validation sets under different experimental settings. All of our experiments are done on an RTX 2080Ti with 11G GPU memory. We run experiments three times and show the best results following (Feng et al., 2021b).

C Other Types of Paraphrase Datasets

To make the input and output carry the same amount of information, one way is to fix D as input and convert utterances into indirect speech as the output. Ganesh and Dingliwal (2019) restructured dialogue into text with complicated rules which are not released and difficult to transfer among datasets under different scenarios. Thus, we only use simple rules to convert all of the utterances into [r_t says,"u_t"] and concatenated as the output. We call this dataset as **DialIndirect**.

Another way is fixing S as output and removing the redundant utterances in D to get the rephrasing input. We take advantage of the idea of oracle extraction for news summarization (Zhou et al., 2018) and regard the combination of dialogue utterances with the highest Rouge scores computed with S as the input. Considering that utterances are

[7] `https://huggingface.co/facebook/bart-large`

Datasets	Input	Output
DialIndirect	$U_{1\sim8}$	Katarina says,“Hello, I got ... we work together” Jill says, “Hi :) ...... nice and sunny”
ExtSum	U_3, U_6	Katarina ...... a flat from Liz. She will ...... after 6 pm.
ExtSumM	$U_{3\sim6}$	Katarina ...... a flat from Liz. She will ...... after 6 pm.
ExtSent/ ExtSentM	U_3	Katarina ...... a flat from Liz.
	U_6	Katarina will ...... after 6 pm.
DSum	$U_{1\sim8}$	Katarina ...... a flat from Liz. She will ...... after 6 pm.
DialSent	$U_{1\sim8}$	Katarina ...... a flat from Liz.
	$U_{1\sim8}$	Katarina will ...... after 6 pm.

Table 8: An illustration of post-training pairs generated from the example in Figure 1. ExtSent and ExtSentM get the same training pairs in this case.

Datasets	Train/Val	IW	OW	CR
SAMSum				
DialIndirect	14,731/818	124.10	157.41	1.31
ExtSum	14,731/818	31.23	23.44	0.94
ExtSumM	14,731/818	66.09	23.44	0.69
EntSent	29,757/1,654	31.05	11.93	0.68
ExtSentM	29,757/1,654	46.45	11.93	0.60
DSum	14,731/818	124.10	23.44	0.25
DialSent	29,757/1,654	149.93	11.93	0.13
DialSumm				
DialIndirect	12,460/500	187.52	215.30	1.16
ExtSum	12,460/500	44.43	30.02	0.84
ExtSumM	12,460/500	94.32	31.02	0.61
EntSent	22,407/840	39.27	17.78	0.65
ExtSentM	22,407/840	61.17	17.78	0.56
DSum	12,460/500	187.52	31.02	0.18
DialSent	22,407/840	214.00	17.78	0.10

Table 9: Statistics of constructed datasets. IW and OW refer to the number of words in the input and output of corresponding dataset. DSum and DialSent are in-list for easier comparison.

highly dependent, we modify the original extraction algorithm by extracting all of the utterances lying between the extracted ones, different from the window-sized snippet selection in (Liu et al., 2021a). Datasets with or without this modification are called **ExtSum** and **ExtSumM** respectively.

A summary S is divided into sentences to construct more rephrase pairs. Similar extraction operations can be done between D and p, and we get **ExtSent** and **ExtSentM** datasets.

An example of the paraphrase pair generated from the dialogue-summary pair in Figure 1 is shown in Table 8. The statistics of post-training datasets derived from SAMSum and DialSumm are shown in Table 9. We compare the performances between different rephrasing approaches with these datasets of our two-stage approach with the fine-tuning-only BART. The results are in Table 10.

Models	Rouge-1	Rouge-2	Rouge-L
SAMSum			
BART	52.06	27.45	48.89
DialIndirect	53.08	28.51	**50.25**
ExtSum	53.20	28.26	49.80
ExtSumM	52.20	27.91	49.74
EntSent	51.82	27.43	49.19
ExtSentM	51.66	27.27	48.96
DSum	52.52	27.51	49.03
DialSent	**53.54**	**28.91**	50.21
DialSumm			
BART	53.01	29.18	51.34
DialIndirect	52.54	29.13	51.68
ExtSum	51.83	27.92	50.33
ExtSumM	52.29	27.72	50.09
EntSent	51.41	27.81	49.65
ExtSentM	52.46	28.86	51.36
DSum	53.27	28.64	51.69
DialSent	**54.73**	**30.47**	**53.46**

Table 10: Comparisons among different post-training approaches and fine-tuning-only BART baseline on dialogue summarization.

DialIndirect performs incredibly well on SAMSum. However, if we use the converted dialogue as input and directly fine-tune the original BART, the results are only 50.91/28.51/50.25 for Rouge-1/2/L. It shows that when accompanied with the post-training stage, the model can learn relationships between speakers and utterances, and boundaries of utterances better than a direct transformation of dialogue inputs. This rule-based transformation falls on DialSumm compared with BART baseline. More complicated rules may lead to better results, but such labored work is not what we are after.

The extraction-based methods fall behind the others. The modification to the algorithm tends to bring more noises than useful information to the input as the results drop mostly. Besides, splitting the summary into sentences doesn't improve the results here. In a word, such hard extractions hurt the intricate discourse and coreference relations among utterances and are not suitable for cross-format data construction.

DialSent with PGG task outperforms other methods and BART consistently across datasets, while DSum with PGG performs almost the same as BART. If we use DialSent data to augment the original DSum during fine-tuning, the results on SAMSum are 44.61/22.81/44.15 for Rouge-1/2/L respectively showing that the data in both datasets is not compatible. Thus, our approach is different from data augmentation. Overall, post-training with cross-format rephrasing intuition does help with dialogue summarization,

D Case Studies

We show more cases as follows.

Dialogue	**Kate**: Hey, do you know if our medical insurance covers hospital costs? **Greg**: Hm, it depends **Mel**: What happened dear? **Kate**: I broke my arm and they're sending me to the hospital :/ **Greg**: Call Linda or ask someone at the reception, they should be able to tell you what kind of package you have **Kate**: thnx
Reference	**Kate** broke her arm and she's going to the hospital. She'd like to know whether her medical insurance covers hospital costs. **Greg** suggests her to call **Linda** or ask someone at the reception about it.
BART	**Kate** broke her arm and they're sending her to the hospital. **Greg** doesn't know if their medical insurance covers hospital costs. **(53.33/37.93/53.19)**
Multi-view	**Kate** broke her arm and they're sending her to the hospital. ***Greg*** *will call* ***Linda*** *or ask someone* at the reception to find out if their insurance covers hospital costs.**(67.64/51.52/56.15)**
DialoBART	**Kate** broke her arm and they're sending her to the hospital . **Greg** advises her to call **Linda** or ask someone at the reception .**(65.57/50.85/67.62)**
DialSent-PGG	**Kate** broke her arm and they're sending her to the hospital. **Greg** advises her to call **Linda** or ask someone at the reception if their insurance covers hospital costs. **(71.64/55.38/62.39)**

Table 11: A case from SAMSum. **Names** are in bold and *unfaithful contents* are in italic. Rouge-1/2/L scores(%) are in parentheses.

The case in Table 11 is a dialogue happened between three speakers from SAMSum. The labeled dialogues, which are directly extracted from Multi-view's and DialoBART's released datasets are shown in Table 12. “|” label for Multi-view refers to the topic transitions and stage transitions for the same dialogue respectively. We can see that topic segments by Multi-view BART are reasonable. However, such linear segmentation is not quite suitable for this dialogue since the first and third topics are the same. “|” in DialoBART just refers to the end of each utterance. DialoBART failed to label any topic transitions or redundant utterances.

Compared to the reference summary, the summary generated by BART lost the information about Greg's suggestion, and DialoBART lost the information about “medical insurance” even though it recognized “medical insurance” as a keyword. Multi-view did incorrect reasoning on who will call Linda. Our model generated a more condensed summary covering the same key points as the reference with the original dialogue as input.

Another case from DialSumm between two speakers is in Table 13. BART recognized “him” in the second utterance as “#Person1#” incorrectly. DialoBART regarded the man as “#Person1#'s friends” which isn't mentioned in the original dialogue. Our model, DialSent-PGG generates a more accurate summary.

<table>
<tr><td>Multi-view Topic</td><td>Kate: Hey, do you know if our medical insurance covers hospital costs? Greg: Hm, it depends | Mel: What happened dear? Kate: I broke my arm and they're sending me to the hospital :/ | Greg: Call Linda or ask someone at the reception, they should be able to tell you what kind of package you have Kate: thnx |</td></tr>
<tr><td>Multi-view Stage</td><td>| Kate: Hey, do you know if our medical insurance covers hospital costs? Greg: Hm, it depends Mel: What happened dear? | Kate: I broke my arm and they're sending me to the hospital :/ | Greg: Call Linda or ask someone at the reception, they should be able to tell you what kind of package you have Kate: thnx</td></tr>
<tr><td>DialoBART</td><td>Kate : Hey , do you know if our medical insurance covers hospital costs ? | Greg : Hm , it depends | Mel : What happened dear ? | Kate : I broke my arm and they're sending me to the hospital | Greg : Call Linda or ask someone at the reception , they should be able to tell you what kind of package you have | Kate : thnx #KEY# Mel Kate Greg Hey do you know if our medical insurance covers hospital costs happened dear Linda reception package</td></tr>
</table>

Table 12: Modified inputs by Multi-view and DialoBART.

Dialogue	**#Person1#**: Like a cat on hot bricks, as you might say. I don ' t believe you are listening at all. **#Person2#**: Sorry, I just worried about him. You know, he should be here an hour ago. **#Person1#**: Don ' t worry him, he has been grown up and I think he can take himself very well. **#Person2#**: But he still does not come back. **#Person1#**: Maybe he is on the way home now.
Reference-1	**#Person2#** is worried about one man, and **#Person1#** thinks that that man might be on the way home now.
Reference-2	**#Person2#** is worried about a man, but **#Person1#** thinks it would be fine.
Reference-3	**#Person2#** is worried about a man but **#Person1#** is not.
BART	***#Person2#*** *is worried about* ***#Person1#*** because he hasn't come back from work. **(43.48/28.57/50.01)**
DialoBART	**#Person2#** is worried about ***#Person1#'s*** *friend* who hasn't come back. **(45.45/30.00/51.87)**
DialSent-PGG	**#Person2#** is worried about a boy who hasn't come back.**(47.62/42.11/53.90)**

Table 13: A case from DialSumm.