Data Augmentation for Low-Resource Dialogue Summarization

Anonymous ACL submission

Abstract

We present *DADS*, a novel **D**ata **A**ugmentation technique for low-resource **D**ialogue **S**ummarization. Our method generates synthetic examples by replacing sections of text from both the input dialogue and summary while preserving the augmented summary to correspond to a viable summary for the augmented dialogue. We utilize pretrained language models that produce highly likely dialogue alternatives while still being free to generate diverse alternatives. We applied our data augmentation method to the SAMSum dataset in low resource scenarios, mimicking real world problems such as chat, thread, and meeting summarization where large scale supervised datasets with human-written summaries are scarce. Through both automatic and human evaluations, we show that DADS shows strong improvements for low resource scenarios while generating topically diverse summaries without introducing additional hallucinations to the summaries.

1 Introduction

As many more language generation tasks are being explored, an outstanding issue is the lack of data available to train generation models. A question that follows is whether it is better to collect and annotate additional data in a particular domain or to generate synthetic data similar to the available data. Considering the elevated cost of collecting data, expertise needed or the difficulty of finding the data, research on data augmentation is warranted. Data augmentation (DA) encompasses methods used to inject additional knowledge into learning systems without explicitly collecting new data; the knowledge injected comes in the form of additional training examples assumed to be silver standard than the collected gold data.

In this paper, we propose an approach for **D**ata **A**ugmentation for **D**ialogue **S**ummarization, aka *DADS*, that creates semantically diverse synthetic examples from a low-resource dataset. Our method modifies both the input dialogue and the target summary while preserving the augmented summary to correspond to a viable summary for the augmented dialogue. First, DADS aligns pairs of utterances from the original dialogue to semantically similar sections in the summary; a large dialogue pretrained model, similar to Meena (Adiwardana et al., 2020), finetuned for dialogue reconstruction, is then used to replace the aligned utterances in the dialogue fabricating new dialogue. A new summary is then synthesized for the newly generated dialogue and the original summary, replacing the aligned sections in the summary using a state-of-the-art pretrained summarization model (Zhang et al., 2019).

Models trained with DADS augmented data produce important performance gains in automated quality metrics for the SAMSum (Gliwa et al., 2019) dialogue summarization dataset in low resource settings, displaying 25% improvement in Rouge when only 10 training examples are available. Gains in performance are present in other low resource settings, such as 50 and 100 examples, but decrease as one would expect as more data is available. As the data augmentation process is inherently noisy, we further investigate whether generation models augmented with DADS are less faithful and analyze other aspects of language generation models such as diversity.

Our main contributions are as follows: (i) We introduce DADS, a novel approach for data augmentation for dialogue summarization for low resource scenarios. (ii) We demonstrate that models trained with DADS augmented data are as faithful as models trained with the original data via human and automated faithfulness metrics. (iii) We found that the outputs generated by DADS augmented models are more diverse than the strong baselines we compare against.

2 Related Work

There is an extensive literature that explores DA for machine learning systems in computer vision (Shorten and Khoshgoftaar, 2019), natural lan-

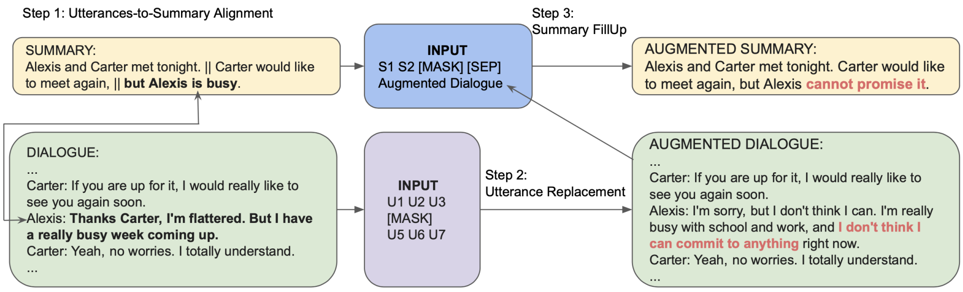

Figure 1: Data augmentation for dialogue summarization. We show how one utterance-summary section pair is aligned (Step 1), here $S3$ and $U4$ are aligned, and replaced both in the input (Step 2) and in the summary (Step 3) producing a new dialogue-summary pair. S's represent sections in the summary and U's utterances in the dialogue.

guage processing (Feng et al., 2021) and other areas. In NLP approaches vary from general-purpose techniques that generate slightly modified copies of existing data; Devries and Taylor (2017) augment examples with noise directly in feature space rather than input space, to domain-specific transformations to create synthetic data, whereas Sennrich et al. (2016) use back-translation to augment text sequences.

Many methods aim to incorporate external knowledge or harness systems and domains where more data is available, e.g. large language models. Recently, Lee et al. (2021) propose example extrapolation by training pretrained language models to extrapolate examples as a few-shot task.

Most prior work in text augmentation has focused on classification tasks, Feng et al. (2021) provide a comprehensive survey of this space. Even though limited, research on data augmentation for language generation has had various approaches to data synthetization, such as corrupting the input text (Xie et al., 2017), the output text (Norouzi et al., 2016) or both (Zhang et al., 2020a). Notably, Schick and Schütze (2021) use pretrained language models and a diverse set of instructions to augment generation datasets in low resource settings, rather than creating training examples.

3 Data Augmentation

We synthesize new training examples by augmenting the dialogue and summary while ensuring that the generated summary is a good abstractive representation for the corresponding dialogue. The augmentation process is done in three steps: utterances-to-summary alignment, dialogue utterance replacement, and summary FillUp. Our workflow is shown in Figure 1 and described below.

Utterances-to-Summary Alignment With the goal of transforming the (dialogue d, summary s) example pairs into a new training example (dialogue d', summary s'), great care has to be taken to avoid them diverging and losing the 'summary-of' relation between the pair. To accomplish this, DADS keeps modifications limited to the aligned sections in the dialogue and summary. Firstly, we align summary spans with utterances in the input. For the particular dataset, SAMSum, summaries are comprised of 1 to 2 sentences. We saw fit to expand the granularity of augmentations to a sub-sentence level by splitting each sentence into clauses. For this, we use an off-the-shelf NLP pipeline annotator spaCy (Honnibal and Montani, 2017).

Next, given the set of all summary clauses and dialogue utterances, we encoded them into a shared space using the universal sentence encoder (Cer et al., 2018). We computed the cosine similarity between the pairs of encoding of utterances and summary clauses. For each clause in the summary, we select the top 20% utterances with the highest similarity scores as our input for augmentation. One (utterances, clause) pair will generate one augmented example.

Dialogue Utterance Replacement We use an auto-regressive encoder-decoder model, inspired by Meena (Adiwardana et al., 2020) and DialogGPT (Zhang et al., 2020b), but initialized from T5-11B (Raffel et al., 2020) and finetuned with a dialog reconstruction loss. The model is trained by randomly masking an utterance from an input example. The task becomes to produce the masked utterance given the masked input example. We use the conversational dataset (SocialMedia), a large-scale high-quality dialog dataset proposed by Meena (Adiwardana et al., 2020) for finetuning. We refer to this finetuned model as DIAL-REPL.

We use DIAL-REPL to generate synthetic alternatives for the selected utterances. Given the original dialogue, the corresponding position of the selected utterance is replaced by a [MASK] token, DIAL-REPL is asked to predict the masked utterance given the input dialogue, the summary and a prompt, as shown in step 2 of Figure 1. We used a standard prompt: "The following conversation is about: " following by the summary and the dialogue. All the selected utterances are replaced one by one in an autoregressive manner: previously generated utterances become part of the input of the next masked position.

Summary FillUp Lastly, we modify the summary by replacing the selected clause with a new one consistent with the augmented synthetic dialogue. We hope this procedure will fulfill two purposes, a more diverse set of summaries, avoiding downstream summarization models to memorize repetitive targets and correct semantic deviations expected to happen during dialogue utterance replacement. We finetuned a large pretrained PEGASUS (Zhang et al., 2019) model for this particular task, to predict a masked sentence in the summary, given the input and summary as context.[1] To generate training data for this model, we converted examples from the CNN/DailyMail (Hermann et al., 2015) dataset by masking a sentence in the gold summary, prepending the masked summary with the input document, separated by a special separator token and tasked the model with predicting the masked sentence, this is akin to the Gap Sentence Generation (Zhang et al., 2019) procedure. For summary augmentation, we mask the summary clause at hand and prepend with the augmented dialogue as input and predict a new replacement clause using the Summary FillUp model.

We augment each annotated dialogue-summary (d, s) pair multiple times, drop duplicated outputs, and keep the rest unique outputs as augmented examples.

4 Experimental Setup

4.1 Low-Resource Dialogue Summarization

We evaluate our method on the SAMSum dialogue summarization dataset (Gliwa et al., 2019), consisting of 14,732, 818 and 819 train, validation and test examples, respectively. To simulate the low-resource summarization setting, we randomly select 10, 50 and 100 annotated examples from the train split for augmentation, then select summarization model parameters with the validation split and report the summarization performance on test split. The inputs and targets were truncated to 1024 and 128.

4.2 Model Comparison

We compare DADS with two other strong baselines: a model trained with no augmented data and a model train using back-translation (Xie et al., 2019) to perturb data instead of language models. We refer to the first model as baseline and the second model as back-translation (Back-trans.) throughout the rest of the paper. In back-translation, we aim to replicate the process we propose of modifying both the dialogue and summary but with a limited semantically-preserving method.[2] For all models, we finetune a large PEGASUS model in two stages: first with the silver standard augmented examples, then we further finetune the model only with the gold examples. For the baseline, we skip the first stage since no silver data is used. The checkpoints are selected using the SAMSum validation split and we report results on the test split. See Table 6 in Appendix for example predictions generated by three models.

4.3 Evaluation Metrics

Along with ROUGE F1 scores (Lin and Hovy, 2003), we report on standard metrics for Semantic Diversity and Faithfulness.

Semantic Diversity We measure word-level semantic diversity in generated summary with the ratio of the number of distinct n-grams and the number of total n-grams. A model that generates semantic-diverse summaries would have a higher proportion of distinct n-grams.

The spikiness of the topic distribution of summaries reflects topic-level diversity. A good summary that captures the main topic in the dialogue would have a sharp topic distribution. A lower entropy value corresponds to a sharper topic distribution. To quantify the spikiness for all the generated summaries, we take the average of the entropy values. Topic distributions of generated summaries are inferred from a MALLET LDA model (McCallum, 2002) trained on the summaries in the SAMSum train split.[3]

Faithfulness Following Maynez et al. (2020a), we report on textual entailment (Pasunuru and Bansal, 2018; Falke et al., 2019; Kryściński et al., 2019) for summary faithfulness evaluation.[4] We also assess faithfulness of generated summaries by human annotation.[5]

5 Results

Compared with the non-augmented baseline, which we call *NoAug*, we find that models trained with

[1] See Appendix A for details about model architecture and parameter selection.

[2] See Appendix B for the back-translation model.

[3] See Appendix C for more details.

[4] See Appendix D for details about the entailment classifier.

[5] See Appendix E for more on the faithfulness assessment.

#Gold Ex	NoAug	Back-translation	DADS
10	25.5/08.3/21.3	28.5/9.6/23.4	**32.5/12.0/27.0**
50	39.8/16.8/32.7	**42.0**/17.9/34.1	41.9/**18.4/34.7**
100	43.0/19.2/35.4	43.2/19.0/35.4	**43.9/19.7/36.1**

Table 1: ROUGE scores (R1/R2/RL) for models trained on 10, 50, and 100 human annotated examples using different data augmentation approaches. For each task we train models in three different sampled sets and report the average score.

Model	#Gold Ex	R1	R2	RL
NoAug	15	29.1	10.5	24.1
NoAug	20	32.4	**12.2**	26.6
DADS	10	**32.5**	12.0	**27.0**
NoAug	60	40.5	17.5	33.6
DADS	50	**41.9**	**18.4**	**34.7**
NoAug	110	43.6	**19.7**	35.9
DADS	100	**43.9**	**19.7**	**36.1**

Table 2: ROUGE scores for DADS models trained with 10, 50 100 number of annotated examples, compared with NoAug baseline models trained with 15, 20, 60 and 110 examples.

Model	Distinct-n		Avg.
	n=1	n=2	Entropy
NoAug	0.162	0.514	6.598
Back-trans.	0.160	0.502	6.604
DADS	**0.176**	**0.581**	**6.597**

Table 3: The number of distinct uni-grams and bi-grams divided by the number of total uni-grams and bi-grams, respectively, higher is better, and average topic distribution entropy, lower is better. All models were trained with 50 annotated examples.

Model	Entail.	Faithfulness	Agree.
Baseline	0.805	2.39	0.66
Back-trans.	0.796	2.41	0.70
DADS	**0.829**	**2.60**	**0.64**

Table 4: Faithfulness assessment (Entailment and Human evaluation) for models trained with 50 annotated examples. Following Durmus et al. (2020), agreement (Agree.) is computed by taking the percentage of the annotators that annotate the majority class for the given (dialogue, summary) pair.

data augmentation generate better quality summaries in terms of ROUGE (see Table 1). Moreover, DADS outperforms the back-translation baseline in all three low resource settings: k = 10, 50, and 100.

For each low resource setting, we investigated the amount of augmented data that achieves the best performance, from 1 to 100 times the amount of gold data. For 10, 50 and 100 examples, back-translation achieves the best performance when 5, 50 and 100 times augmented data is added, respectively. For DADS models, the best performance is achieved when the augmented data amount to 1, 50 and 50 times the gold data, respectively. We hypothesize this is because DADS augmented data contain more diverse and novel information than back-translation augmented examples. Need to notice that $50\times$ DADS augmentation generates about 50% duplicates, resulting in $20\times$ – $30\times$ unique augments.

For each model, the following evaluation and corresponding results are based on the one with the highest ROUGE score in the three runs.

Data Augmentation equivalence to Data Collection. Trying to understand how data augmentation compares with data collection, we set out to find how many additional examples need to be collected to achieve the same performance as DADS augmentation. The result is shown in Table 2. We find that data augmentation when only 100 examples are available is equivalent to more than 10 additionally annotated examples in terms of Rouge-L.

Effect on Semantic Diversity. In Table 3, we show the distinct n-gram proportions and average entropy values for summaries predicted from models trained with 50 annotated examples. Summaries generated by the model with DADS augmentation have the highest proportion of distinct n-grams and the lowest average topic distribution entropy (spikiest topic distribution), suggesting that DADS generates semantically diverse examples. The result also suggests that DADS improved the summarization model's ability to produce textural-diverse, topic-focused summaries.

Effect on Faithfulness. We report the entailment score and the human evaluated faithfulness score in Table 4. We randomly selected 50 documents from the SAMSum test split and assessed the generated summaries from all 3 systems (NoAug, back-translation, and DADS) trained with 50 annotated examples. DADS has the highest Entailment score and faithfulness score. However, through the one-way ANOVA test ($p < 0.01$), we find that differences among all model pairs for both faithfulness are insignificant. This finding suggests that our augmentation approach does not introduce additional hallucinations into the system.

6 Conclusion

We introduced DADS, a new augmentation approach for dialogue summarization tasks. Under 100 annotated examples, the improvement brought from augmentation is roughly equivalent to 10 more annotated examples. Furthermore, we showed that DADS generates semantically diverse synthetic examples. Finally, through automatic and human evaluation, we showed that our augmentation approach does not introduce additional hallucinations to the summarization model.

Ethical Considerations

The nature of text generation leads to multiple ethical considerations when applied to applications. The main failure mode is that the model can learn to mimic target properties in the training data that are not desirable.

Faithfulness and Factuality Since models create new text, there is the danger that they may neither be faithful to the source material nor factual. This can be exacerbated when the data itself has highly abstractive targets, which require the model to generate words not seen in the source material during training. This often leads the model to generate content inconsistent with the source material (Maynez et al., 2020b; Kryscinski et al., 2020; Gabriel et al., 2021).

Trustworthy Data If the data itself is not trustworthy (comes from suspect or malicious sources) the model itself will naturally become untrustworthy as it will ultimately learn the language and topics of the training data. For instance, if the training data is about Obama birther conspiracies, and the model is asked to generate information about the early life of Obama, there is a risk that such false claims will be predicted by the model.

Bias in Data Similarly, biases in the data around gender, race, etc., risk being propagated in the model predictions, which is common for most NLP tasks. This is especially true when the models are trained from non-contemporary data that do not represent current norms and practices (Blodgett et al., 2020).

The above considerations are non-malicious, in that the model is merely learning to behave as its underlying source material. If users of such models are not aware of these issues and do not account for them, e.g., with better data selection, evaluation, etc., then the generated text can be damaging.

Generation models can also be misused in malicious ways. These include generating fake news, spam, and other text meant to mislead large parts of the general population.

References

Daniel De Freitas Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Daniel Matthew Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *ArXiv*, abs/1803.11175.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Terrance Devries and Graham W. Taylor. 2017. Dataset augmentation in feature space. *ArXiv*, abs/1702.05538.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, 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. *arXiv preprint arXiv:1911.12237*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

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.

Kenton Lee, Kelvin Guu, Luheng He, Timothy Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *ArXiv*, abs/2102.01335.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020a. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020b. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Mohammad Norouzi, Samy Bengio, Z. Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In *NIPS*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. *arXiv preprint arXiv:1804.06451*.

Colin Raffel, Noam M. 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. *ArXiv*, abs/1910.10683.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *EMNLP*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *ArXiv*, abs/1511.06709.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Allen Nie, Dan Jurafsky, and A. Ng. 2017. Data noising as smoothing in neural network language models. *ArXiv*, abs/1703.02573.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Rongsheng Zhang, Yinhe Zheng, Jianzhi Shao, Xiao-Xi Mao, Yadong Xi, and Minlie Huang. 2020a. Dialogue distillation: Open-domain dialogue augmentation using unpaired data. *ArXiv*, abs/2009.09427.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020b. Dialogpt : Large-scale generative pre-training for conversational response generation. In *ACL*.

A Summary FillUp Model

Summary FillUp is finetuned from **PEGASUS$_{\mathbf{LARGE}}$** public checkpoint. The model had L = 16, H = 1024, F = 4096, A = 16 (568M parameters), where L denotes the number of layers for encoder and decoder Transformer blocks, H for the hidden size, F for the feed-forward layer size and A for the number of self-attention heads. All finetuning experiments are done with a batch size of 8. For optimization, we use Adafactor (Shazeer and Stern, 2018) with square root learning rate decay with learning rate 0.0001 and a dropout rate of 0.01. The model was decoded with a beam size of 8 and a length-penalty of 0.6.

B Back-translation

For back-translation, we adapted Xie et al. (2019)'s backtranslation implementation to increase diversity. As reported by the authors, the models used are trained in WMT'14 English-French (in both directions). The authors use the hyperparameter *sampling_temp* to control the diversity and quality of the back-translation. We found that setting it to 0.5 yields best augmented examples.

C LDA model

Mallet LDA models are trained with all the 14,732 human annotated summaries in SAMSum train split. We varied the number of topics from 2 to 340, with a step of 2, and select the models with number of topics 100, 200 and 300, the corresponding coherence scores are 0.524, 0.587, and 0.614. Given summaries generated by models trained with DADS and tow baselines, the average topic distribution entropy values calculated from the three LDA models are shown in Table 5. DADS has the lowest average entropy in all three settings.

Model	t=100	t=200	t=300
Baseline	6.598	7.583	8.163
Back-trans.	6.604	7.592	8.172
DADS	**6.597**	**7.583**	**8.162**

Table 5: Average entropy values for Baseline, Back-translation and DADS calculated from three LDA models with number of topics t = 100, 200, and 300.

D Entailment Classifier

Given summary and dialogue, the entailment classifier outputs the probability of the summary entailing the dialogue. We finetuned a transformer-based model, initialized with a pretrained BERT-Large checkpoint (Devlin et al., 2018), on the Multi-NLI dataset (Williams et al., 2017).

E Faithfulness Assessment

We ran a small annotation task with three raters, all proficient in English and NLP reserahcers, who were asked to read the dialogue carefully and then grade the accompanying summary on a scale of 1-4 (fully unfaithful, somewhat unfaithful, somewhat faithful, and fully faithful). A summary is "fully faithful" if all of its content is fully supported or can be inferred from the document.

Gold	Emma was late and missed Andy’s song, but she still had fun.
Dialogue	Emma: Hey it was fun right? George: Yes, certainly.... but why you came so late. you missed andy’s song. Emma: I know :(but still i had a lot of fun. George: yes.. will plan again Emma: yes pleaseeeeee
No Aug. R1/R2/RL	George will plan again for Emma. 16.2 / 9.8 / 16.2
Back Trans. R1/R2/RL	George will come to Emma’s place again. 10.3 / 0.0 / 10.3
DADS R1/R2/RL	Emma came late but still had a lot of fun. George will plan again. 52.2 / 24.0 / 47.8
Gold	Robert wants Fred to send him the address of the music shop as he needs to buy guitar cable.
Dialogue	Robert: Hey give me the address of this music shop you mentioned before Robert: I have to buy guitar cable Fred: < file_other > Fred: Catch it on google maps Robert: thx m8 Fred: ur welcome
No Aug. R1/R2/RL	Robert has to buy guitar cable and Fred has to Catch it on google maps. 40.9 / 29.8 / 40.9
Back Trans. R1/R2/RL	Robert and Fred will meet on google maps. 15.4 / 9.8 / 15.4
DADS R1/R2/RL	Robert wants Fred to give him the address of this music shop. 37.2 / 22.2 / 32.6
Gold	Heidi wants Noah to take items away from the balcony and close all the windows.
Dialogue	Heidi: Could you take the things away from the balcony? I forgot about them and it’s going to rain today. Noah: I’ll do it as soon as I am back home. Heidi: And close all the windows in case of a storm. Noah: of course
No Aug. R1/R2/RL	Noah will take the things away from Heidi’s balcony. 21.3 / 15.4 / 21.3
Back Trans. R1/R2/RL	Noah will take the things away from Heidi. 21.7 / 15.7 / 21.7
DADS R1/R2/RL	Noah will take the things away from the balcony as soon as he is back home. 34.6 / 27.1 / 34.6

Table 6: Dialogue summarization examples: the dialogue, its gold summary and the model generated summaries. We also present the [ROUGE-1, ROUGE-2, ROUGE-L] F1 scores relative to the reference dialogue. The models are trained using 50 annotated examples in SAMSum, with No Augmentation (No Aug.), augmented by Back Translation (Back Trans.), and DADS, respectively.