Abstract

We present DADS, a novel Data Augmentation technique for low-resource Dialogue Summarization. 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 Data Augmentation for Dialogue Summarization, 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-
We synthesize new training examples by augmenting a comprehensive survey of this space. Even producing a new dialogue-summary pair.

to avoid them diverging and losing the ‘summary-

element pairs into a new training example (dia-

Utterances-to-Summary Alignment

We use an auto-regressive encoder-decoder model, inspired by Meena (Adiwardana et al., 2020) and Dialog-

Dialogue Utterance Replacement

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

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-
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.

### 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.
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.

<table>
<thead>
<tr>
<th>#Gold Ex</th>
<th>NoAug</th>
<th>Back-translation</th>
<th>DADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>25.5/8.2/3.1</td>
<td>28.5/9.6/2.3</td>
<td>32.5/12.0/27.0</td>
</tr>
<tr>
<td>50</td>
<td>39.8/16.8/32.7</td>
<td>42.0/17.9/34.1</td>
<td>41.9/18.4/34.7</td>
</tr>
<tr>
<td>100</td>
<td>43.0/19.2/35.4</td>
<td>43.2/19.0/35.4</td>
<td>43.9/19.7/36.1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>#Gold Ex</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoAug</td>
<td>15</td>
<td>29.1</td>
<td>10.5</td>
<td>24.1</td>
</tr>
<tr>
<td>NoAug</td>
<td>20</td>
<td>32.4</td>
<td>12.2</td>
<td>26.6</td>
</tr>
<tr>
<td>DADS</td>
<td>10</td>
<td>32.5</td>
<td>12.0</td>
<td>27.0</td>
</tr>
<tr>
<td>NoAug</td>
<td>60</td>
<td>40.5</td>
<td>17.5</td>
<td>33.6</td>
</tr>
<tr>
<td>DADS</td>
<td>50</td>
<td>41.9</td>
<td>18.4</td>
<td>34.7</td>
</tr>
<tr>
<td>NoAug</td>
<td>110</td>
<td>43.6</td>
<td>19.7</td>
<td>35.9</td>
</tr>
<tr>
<td>DADS</td>
<td>100</td>
<td>43.9</td>
<td>19.7</td>
<td>36.1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Distinct-(n)</th>
<th>Avg. Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoAug</td>
<td>0.162</td>
<td>0.514</td>
</tr>
<tr>
<td>Back-trans.</td>
<td>0.160</td>
<td>0.502</td>
</tr>
<tr>
<td>DADS</td>
<td>0.176</td>
<td>0.581</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Entail.</th>
<th>Faithfulness</th>
<th>Agree.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.805</td>
<td>2.39</td>
<td>0.66</td>
</tr>
<tr>
<td>Back-trans.</td>
<td>0.796</td>
<td>2.41</td>
<td>0.70</td>
</tr>
<tr>
<td>DADS</td>
<td>0.829</td>
<td>2.60</td>
<td>0.64</td>
</tr>
</tbody>
</table>

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


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


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


A Summary FillUp Model

Summary FillUp is finetuned from PEGASUS\textsubscript{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.

<table>
<thead>
<tr>
<th>Model</th>
<th>t=100</th>
<th>t=200</th>
<th>t=300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>6.598</td>
<td>7.583</td>
<td>8.163</td>
</tr>
<tr>
<td>Back-trans.</td>
<td>6.604</td>
<td>7.592</td>
<td>8.172</td>
</tr>
<tr>
<td>DADS</td>
<td><strong>6.597</strong></td>
<td><strong>7.583</strong></td>
<td><strong>8.162</strong></td>
</tr>
</tbody>
</table>

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 researchers, 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 16.2 / 9.8 / 16.2

Back Trans.
R1/R2/RL 10.3 / 0.0 / 10.3

DADS
R1/R2/RL 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 40.9 / 29.8 / 40.9

Back Trans.
R1/R2/RL 15.4 / 9.8 / 15.4

DADS
R1/R2/RL 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 21.3 / 15.4 / 21.3

Back Trans.
R1/R2/RL 21.7 / 15.7 / 21.7

DADS
R1/R2/RL 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.