CONFiT: Toward Faithful Dialogue Summarization with Linguistically-Informed Contrastive Fine-tuning

Anonymous ACL submission

Abstract

Factual inconsistencies in generated summaries severely limit the practical applications of abstractive dialogue summarization. Although significant progress has been achieved by using pre-trained neural language models, substantial amounts of hallucinated content are found during the human evaluation. In this work, we first devised a typology of factual errors to better understand the types of hallucinations generated by current models and conducted human evaluation on popular dialog summarization dataset. We further propose a training strategy that improves the factual consistency and overall quality of summaries via a novel contrastive fine-tuning, called CONFiT. To tackle top factual errors from our annotation, we introduce additional contrastive loss with carefully designed hard negative samples and self-supervised dialogue-specific loss to capture the key information between speakers. We show that our model significantly reduces all kinds of factual errors on both SAMSum dialogue summarization and AMI meeting summarization. On both datasets, we achieve significant improvements over state-of-the-art baselines using both automatic metrics, ROUGE and BARTScore, and human evaluation.

1 Introduction

Text summarization is used to generate a concise and accurate summary of a long text while focusing on the sections that convey the most useful information (Gurevych and Strube, 2004). In recent years, the resurgence of dialogue summarization has attracted significant research attentions (McCowan et al., 2005; Gliwa et al., 2019; Koay et al., 2020; Zhang et al., 2021; Zhong et al., 2021; Zhu et al., 2021; Chen et al., 2021a; Li et al., 2021; Chen et al., 2021c; Fabbri et al., 2021; Chen et al., 2021d). The goal of dialogue summarization is to condense the conversational input into brief sentences version but cover salient information (McCowan et al., 2005; Yuan and Yu, 2020). Significant progress has been made recently on abstractive dialogue summarization with various pre-trained models. However, such pre-trained models are susceptible to generating hallucinate content that is not supported by the source documents (Cao et al., 2018; Maynez et al., 2020; Kryscinski et al., 2020). To tackle the issue of factual inconsistency in dialogue summarization, recent works correctly encode the names of speakers (Zhu et al., 2020), explicitly incorporate coreference information (Liu et al., 2021b), and order the personal named entities (Liu and Chen, 2021). But it is still challenging to improve the quality of summaries generated by different models and decrease the hallucination at the same time.

To better understand the types of hallucinations generated by the pre-trained models, we devised a linguistically motivated taxonomy of factual errors for dialogue summarization, instead of simply classifying the summary as faithful or not. Based on our typology, we defined an annotation protocol for factuality evaluation of dialogue summarization. We then conducted a human evaluation of several pre-trained abstractive summarizers, including BART (Lewis et al., 2020), Pegasus (Zhang et al., 2020), and T5 (Raffel et al., 2020), aiming at identifying the proportion of different types of factual errors and studying the weaknesses of the pre-trained models. Our typology and annotation helps us gain deeper insights into the causes of factual inconsistency. Unlike news summarization (Pagnoni et al., 2021), we found that the challenges posed by dialogue summarization are more related to dialogue flow modeling, informal interactions between speakers, and complex coreference resolution. Figure 1 shows a dialogue-summary pair with three specific errors.

In order to tackle the top factual errors produced by existing models, we propose to replace the most commonly used fine-tuning with a linguistically-informed contrastive fine-tuning approach. For
example, the reason for producing wrong reference errors is that models cannot understand the role in the dialogue, which goes beyond the events. Our goal is to drive the model to pay attention to the grounds of specific errors during the fine-tuning, and learn how to reduce the generation of such errors. To be more specific, CONFiT learns to distinguish whether there are factual errors in the summaries and capture the key information in the dialogue content, such as numbers and person names. Experiments on SAMSUM (Gliwa et al., 2019) and AMI (McCowan et al., 2005) show the generalizability of CONFiT when it is applied to different pre-trained models and datasets. Furthermore, we employ both automatic evaluation and human evaluation on faithfulness and show that CONFiT significantly reduces all different factual errors and generates summaries that are more factually consistent. Moreover, we analytically find that optimizing the contrastive fine-tuning is quite beneficial for improving the robustness of models, which brings further benefits.

Our contributions are as follows:

- We introduce the first typology of factual errors for dialogue summarization and use it to conduct comprehensive annotation and focused analysis.
- Targeting different categories of factual errors in the annotations, we reduce occurrence of such errors generated by various pre-trained models with a novel linguistically-informed contrastive fine-tuning CONFiT approach.
- We validate our method on a widely used dialogue summarization corpus, SAMSUM, and extend it to a meeting summarization corpus AMI. Evaluations of output summaries on automatic metrics like ROUGE, BARTScore as well as human evaluations show that CONFiT outperforms baseline pre-trained models.

2 New Taxonomy of Factuality Errors for Abstractive Dialogue Summarization

In order to gain deeper insights into the types of factuality errors introduced by different abstractive dialogue summarization systems, we proposed a new taxonomy of factuality errors for abstractive dialogue summarization based on our empirical experiments and annotations of the performance of a set of representative baseline summarization models on the SAMSUM dataset, which is a widely-used large-scale dialogue summarization dataset of chat message dialogues in English (see Section 4.1). Specifically, we generate summaries of SAMSUM dialogues using state-of-the-art abstractive dialogue summarization models, including models fine-tuned based on T5 (Raffel et al., 2020), Pegasus (Zhang et al., 2020), BART (Lewis et al., 2020), D-HGN (Xiaochong et al., 2021), and S-BART (Chen and Yang, 2021b). We then manually annotate all different types of errors in these generated summaries that are inconsistent with the source dialogue, compute detailed statistics of all these factuality errors, and then classify them into different categories. Based on our annotation and analysis, we propose a new taxonomy of errors with the majority focusing on factuality error, which includes the following 8 error types:
Category 1 - Missing Information: The content of the generated summary is incomplete compared to the reference.

Example:

[Reference Summary] Williams invites Ms. Blair for a coffee. They will go to her favourite coffee place near the square in a side alley at 2 p.m.

[Model-Generated Summary] Ms. Blair is going to a coffee place near the square in a side alley.

Category 2 - Redundant Information: There is redundant content in the generated summary compared to the reference.

Example:

[Reference Summary] Paula helped Charlotte with correct pronunciation of "Natal Lily."

[Model-Generated Summary] Charlotte asks Paula how to pronounce the name of the plant "Natal Lily." Paula confirms that the stress on the second syllable is 2nd.

Category 3 - Circumstantial Error: Circumstantial information (e.g., date, time, location) about the predicate doesn’t match the reference.

Example:

[Reference Summary] The USA was founded in 1776.

[Model-Generated Summary] The USA was founded in 1767.

Category 4 - Wrong Reference Error: A pronoun is with an incorrect or nonexistent antecedent, or a personal named entity in the generated summary is in the place of a different personal entity in the reference.

Example:


[Model-Generated Summary] Darlene asked Mohit about the test.

Category 5 - Negation Error: This encompasses factual errors resulting from missing or erroneous negation in the generated summary compared to the reference.

Example:


[Model-Generated Summary] Justin does not like books.

Category 6 - Object Error: This covers factual errors resulting from incorrect direct or indirect objects (for non-personal entities only; errors of this nature involving personal entities are designated as Wrong Reference Errors).

Example:

[Reference Summary] Tara raised her glass.

[Model-Generated Summary] Tara raised her spoon.

Category 7 - Tense Error: This encompasses factual errors resulting from discrepancies in grammatical tense between the generated summary and the reference.

Example:

[Reference Summary] The children will go to the library.

[Model-Generated Summary] The children went to the library.

Category 8 - Modality Error: This includes factual errors resulting from modal discrepancies, such getting words like "may", "should", "could" wrong, between the generated summary and the reference.

Example:

[Reference Summary] School may be cancelled today.

[Model-Generated Summary] School is cancelled today.

2.1 Annotation and Analysis

Using our proposed taxonomy of factuality errors, we compute the proportion of each type of factuality errors across different summarization models. We then investigate the model generation behavior that is indicative of errors, which guides the design of our proposed model.

We performed a human evaluation of four model outputs from 19 SAMSum dialogues in order to identify the limitations of abstractive summarization models in dialogue summarization tasks. The four models used in this human evaluation are two BART models with different random seeds.
(ROUGE-L 48 and 49) (Lewis et al., 2020), D-HGN (ROUGE-L 40) (Xiachong et al., 2021), and S-BART (ROUGE-L 48 (Chen and Yang, 2021b)). BART and S-BART are pre-trained models (PLM), and D-HGN is trained from scratch. Since we are focusing on the dialogue domain, most of the factual errors in the model summaries are related to coreference, anaphora, and other dialogue-specific characteristics. In fact, approximately 45% of all errors fall into the categories of Missing Information and Wrong Reference. The distribution of these errors throughout these pre-existing models informs the limitations of each model. Our proposed CONFIT model targets the top errors generated by the current state-of-the-art models to reduce factual inconsistency.

3 CONFIT Model

Standard fine-tuning parameterizes the probability $p_{\alpha}$ of the generator on a task-specific labeled dataset by maximizing cross-entropy loss.

$$\mathcal{L} = - \sum \log P(t_l | t_{<l}, D)$$  \hspace{1cm} (1)$$

However, the cross-entropy loss has several shortcomings that can lead to factual inconsistency in dialogue summarization due to its sub-optimal generalization and instability. We propose a more efficient fine-tuning method CONFIT for factual consistency driven by the intuition that good generalization requires capturing the similarity in one class and contrasting them in other classes. In CONFIT, we introduce two additional losses: contrastive loss and self-supervised loss. We use two weights, actually which is coefficients, to adjust the ratio of $L_{con}$ and $L_{self}$ in the total loss of CONFIT.

The final training objective $\mathcal{J}(\theta)$ of the proposed framework is as follows:

$$\mathcal{J}(\theta) = \mathcal{L} + \alpha \mathcal{L}_{con} + \beta \mathcal{L}_{self}$$  \hspace{1cm} (2)$$

Our linguistically-informed typology and annotation help us gain deeper insights into the causes of different factual errors. To help our models generate more faithful summaries, the proposed CONFIT learns to concentrate on the essential elements of dialogue and capture the dynamic role information as illustrated in Figure 3.

3.1 Contrastive Loss

In order to reduce the occurrence of factual errors, we propose a contrastive loss that uses the following negative sample generation techniques to target each error type in our proposed taxonomy:

- Swap the nouns in the reference summary with each other randomly. This aims to reduce wrong reference and object errors by providing negative samples.

- Swap the verbs in the reference summary with each other randomly. This aims to model reduce circumstance (and, to a lesser extent, tense and modality) errors.

- Mask numbers and years in the dialogue and then pass it into the model to generate a negative sample summary. This aims to reduce circumstance errors.

- Randomly delete 30% of the sentences in the dialogue and then pass it into the model to generate a negative sample summary. This aims to reduce missing information errors.
• Mask-and-fill coreferent entities with BART in the dialogue and then pass it into the model to generate a negative sample summary. This aims to reduce wrong reference errors.

Equation 3 demonstrates our contrastive loss function. During the fine-tuning, we have the positive samples, which is the reference summaries and another set of incorrect summaries, which is the negative samples. The contrastive objectives are learning representations that are invariant to different views of positive pairs; while maximizing the distance between negative pairs (Gunel et al., 2020). Our goal is to maximize the likelihoods of the positive samples and minimize the likelihoods of the negative samples as well. We use the following contrastive learning objective

\[ L_{con} = - \sum_{y_i \neq y_j} \log \frac{\exp(\cos(c_i, c_j))}{\sum_{y_k \neq y_i} \exp(\cos(c_i, c_k))} \]  

(3)

where \( y_i \) and \( y_j \) are positive summary pairs generated by back translation technology and \( y_k \) is from negative set of examples and \( c_i, c_j, c_k \) are their BART encoder representations.

### 3.2 Self-supervised Loss

One unique challenge in abstractive dialogue summarization is the use of first-person pronouns (such as "I" or "we") in speaker utterances, which the model has to correctly identify as being a reference to the speaker. This can lead to wrong reference errors in the summary, as the model cannot understand which participant is speaking and thus cannot accurately resolve first-person references. To address this problem, we design a self-supervised loss that aims to determine whether two tokens belong to the same speaker. Based on these findings, we design a self-supervised loss to enable ConFiT to capture the dynamic roles in the dialogue.

After the BART encoder, the input dialogue is encoded into hidden vectors \( C \). Here, we first randomly select \( k \) pairs of two tokens \( t_m \) and \( t_n \) from the input dialogue, with labels \( s_m \) and \( s_n \) denoting which speaker they are coming from. We also do the same for utterances. Given the concatenation of the encoder representation of dialogue, \( t_m \) and \( t_n \), we use the following loss function to classify whether the two tokens or two utterances are from the same speaker.

\[ L_{self} = - \sum_{m=1}^{k} \sum_{n=1}^{k} \log P(s_m = s_n|t_m, t_n, C) \]  

(4)

This supplementary loss function helps ConFiT keep track of speaker information, thus improving the faithfulness of its summaries for dialogues that contain several first-person references.

### 4 Experiments

#### 4.1 Dataset

We evaluate our new model on the popular SAM-Sum dialogue summarization dataset. Then, we extend our model to meeting summarization with the AMI Meeting Corpus. SAMSum (Gliwa et al., 2019) is a recently proposed large-scale dialogue summarization dataset consisting of 16,369 chat message dialogues in English written by linguists, and each message dialogue is annotated with a multi-sentence summary written by language experts. 75% of the dialogues in the SAMSum dataset (Gliwa et al., 2019) are between two interlocutors, and the other 25% are among three or more interlocutors. The AMI Corpus is another well-known dialogue summarization dataset consisting of 137 multiparty meeting transcripts extracted from 100 hours of meeting recordings. Each meeting transcript in the dataset is also annotated with a generic abstractive summary. We use these two representative dialogue summarization datasets to empirically test our new model’s abstractive summarization performance in the settings of both short conversation-style dialogues and long meeting-style dialogues. See Table 2 for detailed statistics of the two datasets.

#### 4.2 Experiment Settings

In our experiment using SAMSum, we trained BART for 3 epochs with a learning rate of \( 1e-05 \), Pegasus for 20 epochs with a learning rate of \( 1e-04 \), and T5 for 20 epochs with a learning rate of \( 1e-05 \). In our experiment using AMI, we trained BART for 6,000 steps with a learning rate of \( 1e-05 \), Pegasus for 24,000 steps with a learning rate of \( 1e-05 \), and T5 for 20,000 steps with a learning rate of \( 1e-05 \).

#### 4.3 Evaluation Metrics

To evaluate our model, we use three metrics:

**ROUGE** (Lin, 2004): ROUGE measures N-gram overlap between the reference and the automatically generated summaries.

**BARTScore** (Yuan et al., 2021): Because ROUGE scores only measure token overlap, other automated metrics (Rebuffel et al., 2021; Kryscinski et al., 2020; Wang et al., 2020; Scialom et al.,...
<table>
<thead>
<tr>
<th>Model</th>
<th>AMI R-1</th>
<th>AMI R-2</th>
<th>AMI R-L</th>
<th>SAMSum R-1</th>
<th>SAMSum R-2</th>
<th>SAMSum R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extractive and Abstractive Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TextRank (Mihalcea and Tarau, 2004)</td>
<td>35.19*</td>
<td>6.13*</td>
<td>15.70*</td>
<td>29.27*</td>
<td>8.02*</td>
<td>28.78*</td>
</tr>
<tr>
<td>Fast Abs RL (Chen and Bansal, 2018)</td>
<td>38.76</td>
<td>15.13</td>
<td>35.18</td>
<td>40.96</td>
<td>17.18</td>
<td>39.05</td>
</tr>
<tr>
<td>PGN (See et al., 2017)</td>
<td>48.34*</td>
<td>16.02*</td>
<td>23.49*</td>
<td>40.08*</td>
<td>15.28*</td>
<td>36.63*</td>
</tr>
<tr>
<td>PGN(D_{ALL}) (Feng et al., 2021b)</td>
<td>50.91*</td>
<td>17.75*</td>
<td>24.59*</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pre-trained Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5 (Raffel et al., 2020)</td>
<td>42.16</td>
<td>13.94</td>
<td>39.39</td>
<td>48.41</td>
<td>24.79</td>
<td>44.61</td>
</tr>
<tr>
<td>Pegasus (Zhang et al., 2020)</td>
<td>46.02</td>
<td>15.85</td>
<td>43.73</td>
<td>48.04</td>
<td>22.94</td>
<td>43.40</td>
</tr>
<tr>
<td>BART (Lewis et al., 2020)</td>
<td>47.92</td>
<td>16.00</td>
<td>45.36</td>
<td>51.74</td>
<td>26.46</td>
<td>48.72</td>
</tr>
<tr>
<td>Multi-view BART (Chen and Yang, 2020)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5-ConFiT</td>
<td>47.18</td>
<td>13.19</td>
<td>43.55</td>
<td>52.13</td>
<td>27.12</td>
<td>47.62</td>
</tr>
<tr>
<td>Pegasus-ConFiT</td>
<td>48.47</td>
<td>17.61</td>
<td>45.75</td>
<td>52.65</td>
<td>28.21</td>
<td>48.15</td>
</tr>
<tr>
<td>BART-ConFiT</td>
<td>50.31</td>
<td>17.29</td>
<td>47.98</td>
<td>53.89</td>
<td>28.85</td>
<td>49.29</td>
</tr>
</tbody>
</table>

Table 1: Dialogue summarization ROUGE evaluation on the AMI (McCowan et al., 2005) and SAMSum (Gliwa et al., 2019) datasets. We adopt some results reported from the literature (Feng et al., 2021a) and implement the pre-trained models for a fair comparison. All results marked with an asterisk (*) are from Feng et al. (2021b).

### 5 Results

Table 1 shows the ROUGE scores of our models, the baseline models they were fine-tuned from, and a number of other abstractive summarization models on the SAMSum and AMI datasets. Tables 5 and 6 show the average human faithfulness and BART scores respectively for each model’s outputs on 100 SAMSum and 20 AMI dialogues.

We observe that for all three pretrained models ConFiT significantly beat baselines on ROUGE-1, ROUGE-L, and human faithfulness score for both datasets. For BARTScore, we note that, while performance increases on SAMSum for all models, it decreases on AMI. However, given the fact that human evaluators rated the outputs of all three ConFiT models as more faithful than those of their corresponding baselines on both datasets, the decreases in BARTScore on AMI can likely be attributed to the imperfection of automated metrics at capturing faithfulness in text.

#### 5.1 Error Analysis

Tables 3 and 4 show the percentage of summaries that were labeled with each error type in our taxonomy of factual errors (discussed in Section 2.) for both the baseline and ConFiT models on the SAMSum and AMI datasets respectively.

We observe that on SAMSum, our fine-tuning method greatly reduces missing information, redundant information, wrong reference, and circumstance errors for all models. The largest reduction is on the “wrong reference” error type (20%, 7%, and 33% for BART, Pegasus, and T5 respectively).

---

2021) have been proposed to evaluate faithfulness more precisely. BARTScore is a transformer-based measure that scores a dialogue and the corresponding automatically generated summary and has been shown to be strongly correlated with human evaluations of faithfulness (Yuan et al., 2021).

**Human Evaluation:** Finally, we conduct human evaluations on 100 SAMSum (Gliwa et al., 2019) and 20 AMI (McCowan et al., 2005) dialogues. Tang et al. (2021) found that Likert scales are a more consistent measure of factuality for abstractive dialogue summarization than Best-Worst Scaling. We have human evaluators directly rate the summaries on a scale from 1 to 10 corresponding to their faithfulness. In addition, using the error taxonomy proposed in Section 2, we have them mark whether each error type appeared in the given summary. We do this in a blinded fashion, so that the annotators do not see the corresponding model of the summary. Additionally, in order to prevent model information from leaking to the annotators, we randomly shuffle outputs within each dialogue before assigning them to annotators.
<table>
<thead>
<tr>
<th>Error Type</th>
<th>BART</th>
<th>BART-ConFiT</th>
<th>Pegasus</th>
<th>Pegasus-ConFiT</th>
<th>T5</th>
<th>T5-ConFiT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Information</td>
<td>55%</td>
<td>44%</td>
<td>56%</td>
<td>50%</td>
<td>63%</td>
<td>48%</td>
</tr>
<tr>
<td>Redundant Information</td>
<td>12%</td>
<td>7%</td>
<td>7%</td>
<td>4%</td>
<td>7%</td>
<td>4%</td>
</tr>
<tr>
<td>Wrong Reference</td>
<td>37%</td>
<td>17%</td>
<td>25%</td>
<td>18%</td>
<td>46%</td>
<td>13%</td>
</tr>
<tr>
<td>Circumstance</td>
<td>14%</td>
<td>8%</td>
<td>16%</td>
<td>10%</td>
<td>8%</td>
<td>9%</td>
</tr>
<tr>
<td>Negation</td>
<td>4%</td>
<td>1%</td>
<td>7%</td>
<td>2%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Object</td>
<td>10%</td>
<td>6%</td>
<td>4%</td>
<td>7%</td>
<td>2%</td>
<td>7%</td>
</tr>
<tr>
<td>Tense</td>
<td>2%</td>
<td>1%</td>
<td>3%</td>
<td>1%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Modality</td>
<td>6%</td>
<td>1%</td>
<td>3%</td>
<td>5%</td>
<td>5%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Table 3: Percentage of autogenerated summaries containing each error type, according to our human evaluation of model outputs from 100 SAMSum dialogues. Note that a single summary can contain multiple error types, so they do not add up to 100%.

<table>
<thead>
<tr>
<th>Error Type</th>
<th>BART</th>
<th>BART-ConFiT</th>
<th>Pegasus</th>
<th>Pegasus-ConFiT</th>
<th>T5</th>
<th>T5-ConFiT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Information</td>
<td>90%</td>
<td>85%</td>
<td>80%</td>
<td>70%</td>
<td>80%</td>
<td>85%</td>
</tr>
<tr>
<td>Redundant Information</td>
<td>10%</td>
<td>15%</td>
<td>60%</td>
<td>25%</td>
<td>0%</td>
<td>25%</td>
</tr>
<tr>
<td>Wrong Reference</td>
<td>35%</td>
<td>30%</td>
<td>35%</td>
<td>30%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Circumstance</td>
<td>35%</td>
<td>35%</td>
<td>30%</td>
<td>30%</td>
<td>40%</td>
<td>35%</td>
</tr>
<tr>
<td>Negation</td>
<td>20%</td>
<td>15%</td>
<td>5%</td>
<td>15%</td>
<td>25%</td>
<td>0%</td>
</tr>
<tr>
<td>Object</td>
<td>45%</td>
<td>40%</td>
<td>45%</td>
<td>25%</td>
<td>55%</td>
<td>55%</td>
</tr>
<tr>
<td>Tense</td>
<td>10%</td>
<td>10%</td>
<td>0%</td>
<td>5%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>Modality</td>
<td>10%</td>
<td>15%</td>
<td>5%</td>
<td>5%</td>
<td>20%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Table 4: Percentage of autogenerated summaries containing each error type, according to our human evaluation of model outputs from 20 AMI dialogues. Note that a single summary can contain multiple error types, so they do not add up to 100%.

<table>
<thead>
<tr>
<th>Faithfulness Score</th>
<th>SAMSum</th>
<th>AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART</td>
<td>5.540</td>
<td>4.850</td>
</tr>
<tr>
<td>BART-ConFiT</td>
<td>7.250</td>
<td>5.600</td>
</tr>
<tr>
<td>Pegasus</td>
<td>6.260</td>
<td>5.250</td>
</tr>
<tr>
<td>Pegasus-ConFiT</td>
<td>6.770</td>
<td>5.895</td>
</tr>
<tr>
<td>T5</td>
<td>5.422</td>
<td>4.150</td>
</tr>
<tr>
<td>T5-ConFiT</td>
<td>6.920</td>
<td>4.950</td>
</tr>
</tbody>
</table>

Table 5: Average faithfulness score (on a scale of 1-10) given to each model by human evaluators on 100 SAMSum and 20 AMI dialogues. Highest scores for each dataset have been bolded.

<table>
<thead>
<tr>
<th>BARTScore</th>
<th>SAMSum</th>
<th>AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART</td>
<td>-1.613</td>
<td>-3.644</td>
</tr>
<tr>
<td>BART-ConFiT</td>
<td>-1.468</td>
<td>-3.669</td>
</tr>
<tr>
<td>Pegasus</td>
<td>-1.615</td>
<td>-2.967</td>
</tr>
<tr>
<td>Pegasus-ConFiT</td>
<td>-1.608</td>
<td>-3.369</td>
</tr>
<tr>
<td>T5</td>
<td>-1.993</td>
<td>-3.406</td>
</tr>
<tr>
<td>T5-ConFiT</td>
<td>-1.677</td>
<td>-3.798</td>
</tr>
</tbody>
</table>

Table 6: Average BARTScore for each model on 100 SAMSum and 20 AMI dialogues. Highest scores for each dataset have been bolded.

likely owing to the self-supervised loss function introduced in Section 3.2 that was designed to help the model more effectively capture speaker information. For AMI, however, our fine-tuning method is not as consistent at reducing the frequency of each error type across models. It is possible that this is due to sample size (20 AMI dialogues vs. 100 SAMSum dialogues).

5.2 Case Study

Figure 4 shows the results of human annotation on the model outputs of a selected SAMSum dialogue. Note that all of the autogenerated summaries, both baseline and ConFiT, were marked as having missing information errors by the annotator, likely due to the omission of Ernest’s relief upon hearing that the car that was crashed into did not belong to Mike. As a result, none of the models achieved a perfect factuality score on this dialogue; however, the scores for each ConFiT model were higher than those of their corresponding baselines.

It can be observed that while baseline BART outputs a summary with a circumstance error, mistakenly asserting that Mike parked his car on Ernest’s street, the BART+ConFiT fixes this error, correctly asserting that Mike took his car to the garage today; as a result, the human annotator gave this summary a higher score than the predicted summary from baseline BART. Baseline T5 outputs a summary with two coreference errors; specifically, it contains a missing subject in the first sentence and incorrectly implies that the car that got crashed into belonged to Mike in the second sentence. The T5+ConFiT is able to fix both of these errors,
adding "Mike" to the beginning of the first sentence and changing "his red Honda" to "a red Honda just like Mike's" in the second sentence. Similarly, the output of baseline Pegasus contains a coreference error in the first sentence, implying that Mike owns the car that was crashed into while the output of Pegasus+CONFiT does not.

### 6 Related Work

Multi-party dialogues are especially challenging to summarize using automated models, given that they often contain pauses, false starts, reconfirmations, hesitations, and speaker interruptions (Sacks et al., 1978; Feng et al., 2021a; Chen and Yang, 2021a). Previous work in the field has addressed these challenges by incorporating semantic features, including keywords (Zhu et al., 2020), domain terminologies (Koay et al., 2020), topics (Zhao et al., 2020; Liu et al., 2021a), entailment knowledge (Li et al., 2018), and background knowledge (Feng et al., 2021c). Other works exploit personal named entities (Liu and Chen, 2021) and coreference information (Liu et al., 2021b) to learning to distinguish complex coreferent relationships expressed through personal pronouns (including the first person "I") in the conversation (Lei et al., 2021). Researchers have also explored conversational structure (Zhao et al., 2021), utterance flow modelling (Chen et al., 2021b), syntactic structure (Lee et al., 2021), granularity control (Wu et al., 2021), but they have not yet converged to a simple and practical solution.

Our proposed taxonomy of factual errors and annotations help us gain deeper insights into the causes of factual inconsistency in abstractive dialogue summarization outputs.

### 7 Conclusion

We presented CONFiT, a novel method to improve the faithfulness of abstractive dialogue summarization models via contrastive and self-supervised fine-tuning. By adapting the objective function during fine-tuning to incorporate a contrastive loss that learns to distinguish positives from examples with factual errors, and a self-supervised dialogue-specific loss that captures important dialogue information flow between multiple interlocutors, CONFiT can significantly improve the faithfulness of the abstractive summaries generated by transformer-based sequence-to-sequence language models, and reduce multiple categories of factuality errors in the abstractive summaries by large margins. In our experiment on SAMSum and AMI, we demonstrated that CONFiT achieves better empirical performance compared to the baseline models fine-tuned with the traditional cross-entropy loss, based on both automatic evaluation metrics and human evaluation. Our work provides new insights into improving the faithfulness of abstractive summarization systems using carefully designed novel objective functions for fine-tuning that captures important structures and features of the text to summarize.
8 Ethics Statement

Human Evaluation We recruited seven volunteer participants, requesting speakers of English. These annotators are participating voluntarily. Our participants are free to opt out of the study at any point in time. We have written four scripts for use in the annotation process: (1) the first script generates an annotation spreadsheet and a key spreadsheet from the model outputs. The annotation spreadsheet does not contain the model names; however, it contains an id that can be used to recover the model name from the key spreadsheet. For ease of annotation, summaries from the same dialogue are grouped together; however, they are randomly shuffled within each dialogue so that the annotators cannot guess from the ordering as to which model is which. (2) The second script splits an annotation spreadsheet into multiple spreadsheets so that the work can be distributed amongst annotators. (3) The third one merges these spreadsheets back together after the annotation process is finished. (4) The last script recovers the model names from the key spreadsheet and inserts them into the annotation spreadsheet. Each evaluator is asked to examine whether there is an error and the full context (dialogue, generated summaries, and reference) and give a score on a scale of 1 to 10 for each of the criteria. We only consider faithfulness, instead of general quality. E.g. 1: very poor, 3: poor, 5: neutral; 7: good; 10: very good. We asked each internal annotator to evaluate 300 samples.

Other Ethical Issues (1) We did not use any personally identifiable information in the experiments. (2) The goal of the project, improving the faithfulness of automatically generated summaries, is to make the output of the summarization system more reliable and minimize confusion for the readers of the summaries. (3) We used existing summarization datasets that do not contain any sensitive information and are unlikely to cause any harm to the annotators.

References


