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

Dialogue (Copy 1)	
Hannah:	Hey, do you have Betty's number? Amanda: Lemme check
Amanda:	Sorry, can't find it. Ask Larry. He called her last time we were at the park together.
Hannah:	I don't know him well.
Amanda:	Don't be shy, he's very nice.
Hannah:	If you say so… I'd rather you texted him.
Amanda:	Okay. I just texted him.
Hannah:	Urgh.. Alright. Bye.

(a) Coreference Error

Dialogue (Copy 2)	
Hannah:	Hey, do you have Betty's number? Amanda: Lemme check
Amanda:	Sorry, can't find it. Ask Larry. He called her last time we were at the park together.
Hannah:	I don't know him well.
Amanda:	Don't be shy, he's very nice.
Hannah:	If you say so… I'd rather you texted him.
Amanda:	Okay. I just texted him.
Hannah:	Urgh.. Alright. Bye.

(b) Modality & Tense Error

Dialogue (Copy 3)	
Hannah:	Hey, do you have Betty's number? Amanda: Lemme check
Amanda:	Sorry, can't find it. Ask Larry. He called her last time we were at the park together.
Hannah:	I don't know him well.
Amanda:	Don't be shy, he's very nice.
Hannah:	If you say so… I'd rather you texted him.
Amanda:	Okay. I just texted him.
Hannah:	Urgh.. Alright. Bye.

(c) Missing Information

Reference
Hannah needs Betty's number but Amanda doesn't have it. Amanda needs to contact Larry.

Generated Summary
Amanda can't find Betty's number. Larry called her last time they were at the park. Amanda will text Larry.

Figure 1: Sample summary of a SAMSum dialogue (Gliwa et al., 2019). The summary is generated by BART (Lewis et al., 2020). Errors are highlighted.

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 (Xiachong 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:

[Reference Summary] *Mohit asked Darlene about the test.*

[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:

[Reference Summary] *Justin likes books.*

[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

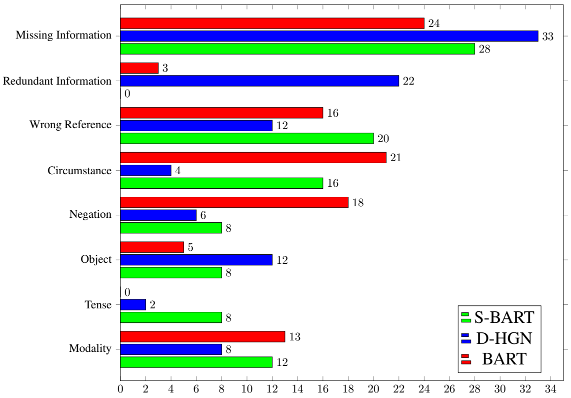


Figure 2: Percentage of error types in each model during preliminary human evaluation of 19 SAMSum dialogues.

(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

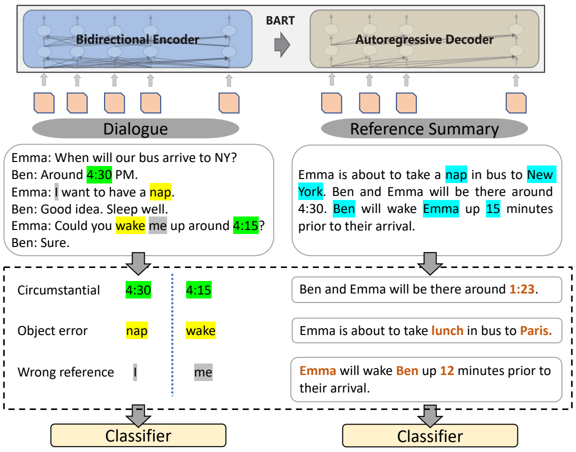
Standard fine-tuning parameterizes the probability p_α of the generator on a task-specific labeled dataset by maximizing cross-entropy loss.

$$\mathcal{L} = -\sum \log P(\tilde{t}_l | t_{<l}, \mathbf{D}) \qquad (1)$$

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:

$$\mathcal{J}(\theta) = \mathcal{L} + \alpha\mathcal{L}_{\text{con}} + \beta\mathcal{L}_{\text{self}} \qquad (2)$$

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.



Bidirectional Encoder
BART
Autoregressive Decoder
Dialogue
Reference Summary
Emma: When will our bus arrive to NY?
Ben: Around 4:30 PM.
Emma: I want to have a nap.
Ben: Good idea. Sleep well.
Emma: Could you wake me up around 4:15?
Ben: Sure.
Emma is about to take a nap in bus to New York. Ben and Emma will be there around 4:30. Ben will wake Emma up 15 minutes prior to their arrival.
Circumstantial
4:30
4:15
Object error
nap
wake
Wrong reference
I
me
Ben and Emma will be there around 1:23.
Emma is about to take lunch in bus to Paris.
Emma will wake Ben up 12 minutes prior to their arrival.
Classifier
Classifier

Figure 3: A demonstration of our model.

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 the 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

$$\mathcal{L}_{con} = -\sum_{y_j \neq y_i} \log \frac{\exp(\cos(\boldsymbol{c}_i, \boldsymbol{c}_j))}{\sum\limits_{y_k \neq y_i} \exp(\cos(\boldsymbol{c}_i, \boldsymbol{c}_k))} \quad (3)$$

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 $\boldsymbol{c}_i$,$\boldsymbol{c}_j$, $\boldsymbol{c}_k$ 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.

$$\mathcal{L}_{self} = -\sum_{m=1}^{k}\sum_{n=1}^{k} \log P(s_m = s_n | t_m, t_n, C) \quad (4)$$

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 SAMSum 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.,

Model	AMI			SAMSum		
	R-1	R-2	R-L	R-1	R-2	R-L
Extractive and Abstractive Models						
TextRank (Mihalcea and Tarau, 2004)	35.19*	6.13*	15.70*	29.27*	8.02*	28.78*
Fast Abs RL (Chen and Bansal, 2018)	38.76	15.13	35.18	40.96	17.18	39.05
PGN (See et al., 2017)	48.34*	16.02*	23.49*	40.08*	15.28*	36.63*
PGN($\mathcal{D}_{\text{ALL}}$) (Feng et al., 2021b)	50.91*	17.75*	24.59*	-	-	-
Pre-trained Models						
T5 (Raffel et al., 2020)	42.16	13.94	39.39	48.41	24.79	44.61
Pegasus (Zhang et al., 2020)	46.02	15.85	43.73	48.04	22.94	43.40
BART (Lewis et al., 2020)	47.92	16.00	45.36	51.74	26.46	48.72
Multi-view BART (Chen and Yang, 2020)	-	-	-	49.52	26.52	48.29
Ours						
T5-ConFiT	47.18	13.19	43.55	52.13	27.12	47.62
Pegasus-ConFiT	48.47	17.61	45.75	52.65	28.21	48.15
BART-ConFiT	50.31	17.29	47.98	53.89	28.85	49.29

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).

	Dialogue	*Speakers*	*Turns*	*Length*
SAMSum				
Train	14732	2.40	11.17	23.44
Validation	818	2.39	10.83	23.42
Test	819	2.36	11.25	23.12
AMI				
All	137	4	289	322

Table 2: Details about SAMSum and AMI.

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.

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),

Error Type	BART	BART-ConFiT	Pegasus	Pegasus-ConFiT	T5	T5-ConFiT
Missing Information	55%	44%	56%	50%	63%	48%
Redundant Information	12%	7%	7%	4%	7%	4%
Wrong Reference	37%	17%	25%	18%	46%	13%
Circumstance	14%	8%	16%	10%	8%	9%
Negation	4%	1%	7%	2%	1%	1%
Object	10%	6%	4%	7%	2%	7%
Tense	2%	1%	3%	1%	2%	2%
Modality	6%	1%	3%	5%	5%	8%

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%.

Error Type	BART	BART-ConFiT	Pegasus	Pegasus-ConFiT	T5	T5-ConFiT
Missing Information	90%	85%	80%	70%	80%	85%
Redundant Information	10%	15%	60%	25%	0%	25%
Wrong Reference	35%	30%	35%	30%	50%	50%
Circumstance	35%	35%	30%	30%	40%	35%
Negation	20%	15%	5%	15%	25%	0%
Object	45%	40%	45%	25%	55%	55%
Tense	10%	10%	0%	5%	10%	10%
Modality	10%	15%	5%	5%	20%	10%

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%.

Faithfulness Score	SAMSum	AMI
BART	5.540	4.850
BART-ConFiT	**7.250**	**5.600**
Pegasus	6.260	5.250
Pegasus-ConFiT	**6.770**	**5.895**
T5	5.422	4.150
T5-ConFiT	**6.920**	**4.950**

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.

BARTScore	SAMSum	AMI
BART	-1.613	**-3.644**
BART-ConFiT	**-1.468**	-3.669
Pegasus	-1.615	**-2.967**
Pegasus-ConFiT	**-1.608**	-3.369
T5	-1.993	**-3.406**
T5-ConFiT	**-1.677**	-3.798

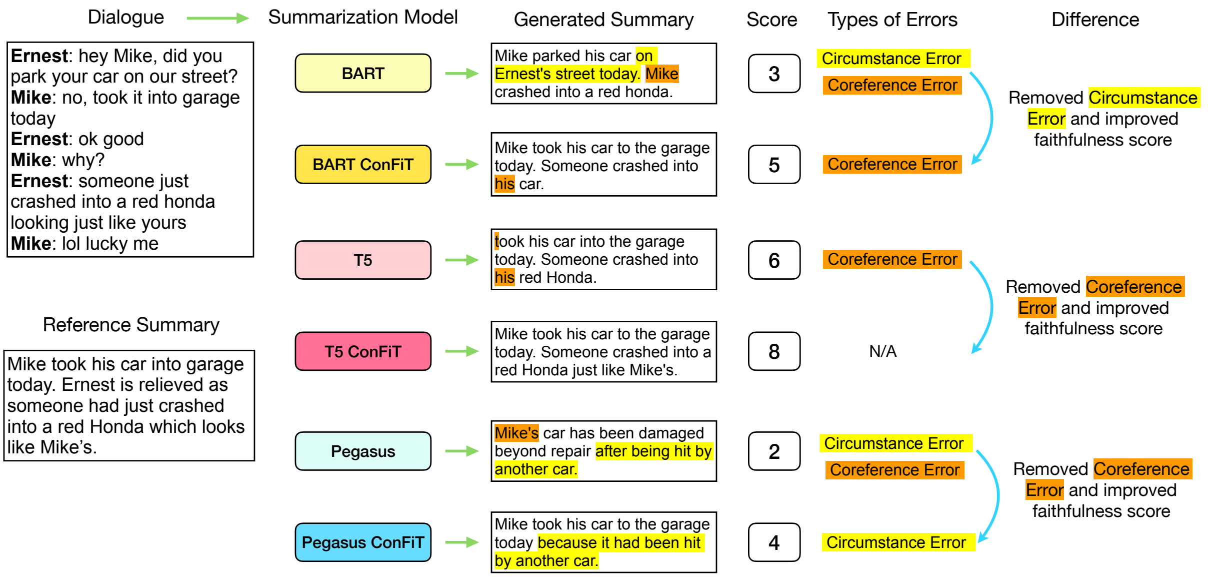
Table 6: Average BARTScore for each model on 100 SAMSum and 20 AMI dialogues. Highest scores for each dataset have been bolded.

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,



Ernest: hey Mike, did you park your car on our street?
Mike: no, took it into garage today
Ernest: ok good
Mike: why?
Ernest: someone just crashed into a red honda looking just like yours
Mike: lol lucky me

Mike took his car into garage today. Ernest is relieved as someone had just crashed into a red Honda which looks like Mike's.

Mike parked his car on Ernest's street today. Mike crashed into a red honda.

Mike took his car to the garage today. Someone crashed into his car.

took his car into the garage today. Someone crashed into his red Honda.

Mike took his car to the garage today. Someone crashed into a red Honda just like Mike's.

Mike's car has been damaged beyond repair after being hit by another car.

Mike took his car to the garage today because it had been hit by another car.

Figure 4: Model outputs for selected SAMSum dialogue, along with the corresponding reference summary, human factuality scores, and 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

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

Jiaao Chen and Diyi Yang. 2021a. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021b. 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, Online. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021a. Summscreen: A dataset for abstractive screenplay summarization.

Wang Chen, Piji Li, Hou Pong Chan, and Irwin King. 2021b. Dialogue summarization with supporting utterance flow modelling and fact regularization. *Knowledge-Based Systems*, 229:107328.

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, Melbourne, Australia. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021c. 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.

Yulong Chen, Yang Liu, and Yue Zhang. 2021d. Dialogsum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313.

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. *arXiv preprint arXiv:2105.12544*.

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

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770, Geneva, Switzerland. COLING.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5689–5695, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Seolhwa Lee, Kisu Yang, Chanjun Park, João Sedoc, and Heuiseok Lim. 2021. Who says like a style of vitamin: Towards syntax-aware dialoguesummarization using multi-task learning. *arXiv preprint arXiv:2109.14199*.

Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. 2021. Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7823–7827. IEEE.

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

Daniel Li, Thomas Chen, Albert Tung, and Lydia Chilton. 2021. Hierarchical summarization for long-form spoken dialog.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021a. Topic-aware contrastive learning for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Zhengyuan Liu and Nancy F. Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the Association for Computational Linguistics: EMNLP 2020*, Online. 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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. 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.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas

Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proc. of EMNLP*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

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

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-questeval: A referenceless metric for data to text semantic evaluation. *arXiv preprint arXiv:2104.07555*.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Xiangru Tang, Alexander R. Fabbri, Ziming Mao, Griffin Adams, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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

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

Lin Yuan and Zhou Yu. 2020. Abstractive dialog summarization with semantic scaffolds.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*.

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

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. EmailSum: Abstractive email thread 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 6895–6909, Online. Association for Computational Linguistics.

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

Lulu Zhao, Zeyuan Yang, Weiran Xu, Sheng Gao, and Jun Guo. 2021. Improving abstractive dialogue summarization with conversational structure and factual knowledge.

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

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

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