ConFiT: Contrastive Fine-Tuning for Abstractive Dialogue Summarization

Xiangru Tang[1], Borui Wang[1], Arjun Nair[1], Bingyao Wang[1], Jai Desai[1], Aaron Wade[1], Haoran Li[2], Asli Celikyilmaz[2], Yashar Mehdad[2], Dragomir Radev[1]

[1]Yale University [2]Facebook AI

LILY Lab

Introduction

Recently, various neural encoder-decoder models pioneered by the Seq2Seq framework have been proposed to achieve the goal of generating more abstractive summaries by learning to map input text to output text. At a high level, such neural models can freely generate summaries without any constraint on the words or phrases used. Moreover, their format is closer to human-edited summaries and output is more readable and fluent. However, the neural model's abstraction ability is a double-edged sword. A commonly observed problem with the generated summaries is the distortion or fabrication of factual information in the article.

However, existing neural abstractive summarization models are highly prone to generate factual inconsistency errors. It refers to the phenomenon that the summary sometimes distorts or fabricates the facts in the article. Recent studies show that up to 30% of the summaries generated by abstractive models contain such factual inconsistencies. This brings serious problems to the credibility and usability of abstractive summarization systems.

Materials and Methods

In this work, we provide a typology of factual errors to highlight the types of errors and move away from a binary understanding of factuality.

We further propose a training strategy that improves the factual consistency and overall quality of summaries via a novel contrastive fine-tuning, called ConFiT.

Specifically, we employ different modular objectives that each target a specific type of error: (1) contrastive loss, which utilizes hard negative samples that are factually inconsistent (2) self-supervised dialogue-specific loss, which captures dialogue information flow between several participants.

Targeting different categories of factual errors in the annotations, we propose a contrastive fine-tuning approach ConFiT to minimize the rate of occurrence of such errors. We validate our method on two widely used dialogue summarization datasets, SAMSum and AMI. \baby outperforms large pre-trained language models on factual consistency metrics and human evaluation.

Model	AMI R-1	AMI R-2	AMI R-L	SAMSum R-1	SAMSum R-2	SAMSum R-L
Extractive and Abstractive Models						
TextRank (Mihalcea and Tarau, 2004)	35.19	6.13	15.70	30.72	4.69	12.97
Fast Abs RL (Chen and Bansal, 2018)	38.76	15.13	35.18	40.96	17.18	39.05
PGN (See et al., 2017)	42.60	14.01	22.62	40.08	15.28	36.63
PGN($\mathcal{D}_{\mathrm{ALL}}$) (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. ROUGE scores for baseline and ConFiT models.

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 2. Human factuality scores for baseline and ConFit.

BART Score	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 3. BART scores for baseline and ConFit models.

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 4. Percentage of autogenerated summaries containing each error type, according to our human evaluation of model outputs from 100 SAMSum dialogues.

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 5. Percentage of autogenerated summaries containing each error type, according to our human evaluation of model outputs from 20 AMI dialogues.

Results

We validate our method on two widely used dialogue summarization datasets, SAMSum and AMI. ConFiT outperforms large pre-trained language models on factual consistency metrics and human evaluation. ConFiT also achieves better results on general quality, like ROUGE.

We observe that all three \baby models perform better than their corresponding baselines on ROUGE-1, ROUGE-L, and human factuality on 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 \baby models as more factual 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 factuality in text.

Conclusion

In this paper we presented ConFiT, a novel method to improve the faithfulness of abstractive dialogue summarization models through the approach of contrastive and self-supervised fine-tuning. By adapting the objective function during fine-tuning to incorporate a contrastive loss that learns to distinguish false positives from factual errors, and a self-supervised dialogue-specific loss that captures important dialogue information flow between multiple interlocutors, our new method is able to 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 the SAMSum dataset and the AMI Meeting Corpus, we demonstrated that our proposed ConFiT method achieves better empirical performance compared to the baseline models fine-tuned with the traditional cross-entropy loss in both automatic evaluation metrics and human evaluation metrics. 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.