ConFiT: Contrastive Fine-Tuning for Abstractive Dialogue Summarization

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

Yale University[1], Facebook AI[2]

LILY Lab

Introduction

Dialogue summarization is the task of using an automated model to produce a concise, acceptable summary of a piece of dialogue. **Abstractive** dialogue summarization, in particular, involves the generation of new content in order to create the summary, as opposed to **extractive** methods which simply pull the most important sentences and/or phrases directly from the text to form its output.

A common metric for measuring the efficacy of abstractive dialogue summarization systems is the **ROUGE** score, which compares the n-gram overlap between reference and autogenerated summaries. While ROUGE does have some correlation with human faithfulness, it is far from a perfect proxy. Thus, hyperoptimizing for ROUGE scores alone, without making considerations for the actual faithfulness of the output may cause state-of-the-art models in the field to veer further and further away from true faithfulness.

Methods

Our team has developed a contrastive fine-tuning method called **ConFiT**, which utilizes the following supplementary loss functions to improve the faithfulness of existing abstractive dialogue summarization models:

Contrastive Loss: Uses negative samples (incorrect summaries) to teach the model what *not* to output. These negative samples are generated by mutating the dialogue in order to induce an error and then passing that mutated dialogue through a summarization model.

Self-Supervised Loss: Helps the model determine whether two tokens belong to the same speaker.

We evaluated our model's performance using two automated metrics - ROUGE and BARTScore. In addition, we had six human evaluators conduct a blinded annotation of model outputs from 100 SAMSum and 20 AMI dialogues in which they marked the errors that appeared in a given summary and assigned a factuality score from 1-10.

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($\mathcal{D}_{\text{ALL}}$) (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. 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 models.

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

Our models achieve state-of-the-art performance on SAMSum and AMI, outperforming the baselines on ROUGE-1, ROUGE-L, and human factuality scores. With BARTScore, however, our models show increased performance on SAMSum and decreased performance on AMI, perhaps owing to the imperfection of automated metrics at capturing human factuality.

On SAMSum, ConFiT greatly decreases the occurrence of missing information, redundant information, wrong reference, and circumstance errors for all models. The largest decreases can be seen for the wrong reference error type (20%, 7%, and 33% for BART, Pegasus, and T5 respectively), likely owing to the self-supervised loss function that was designed to reduce such errors. For AMI, however, ConFiT is not as consistent at reducing the frequency of each error type, with increases and decreases in percentage varying largely from model to model. It is possible that this is due to sample size and that the results may be more consistent if we were to evaluate over a larger set of AMI dialogues (100, for example).

Conclusion

Our team has developed ConFiT, a contrastive fine-tuning approach that leads to state-of-the-art performance on both SAMSum and AMI as judged by human evaluators and automated metrics such as ROUGE and BARTScore. By utilizing negative sample generation techniques that explicitly induce missing information errors and a supplementary loss function that incorporates speaker information, we were able to greatly improve the performance of BART, Pegasus, and T5. Our work paves the way for future growth in the field of abstractive dialogue summarization by providing a new taxonomy of errors for this task and highlighting the importance of utilizing speaker information and negative sample generation techniques that focus on inducing specific errors in this taxonomy.