Human-in-the-loop Abstractive Dialogue Summarization

Jiaao Chen[†] **Mohan Dodda**[†] **Diyi Yang**[◇]
[†] Georgia Institute of Technology [◇] Stanford University
{jchen896,mohandodda}@gatech.edu
diyiy@cs.stanford.edu

Abstract

Abstractive dialogue summarization has received increasing attention recently. Despite the fact that most of the current dialogue summarization systems are trained to maximize the likelihood of human-written summaries and have achieved significant results, there is still a huge gap in generating high-quality summaries as determined by humans, such as coherence and faithfulness, partly due to the misalignment in maximizing a single human-written summary. To this end, we propose to incorporate different levels of human feedback into the training process. This will enable us to guide the models to capture the behaviors humans care about for summaries. Specifically, we ask humans to highlight the salient information to be included in summaries to provide the *local feedback*, and to make overall comparisons among summaries in terms of coherence, accuracy, coverage, concise and overall quality, as the *global feedback*. We then combine both local and global feedback to fine-tune the dialog summarization policy with Reinforcement Learning. Experiments conducted on multiple datasets demonstrate the effectiveness and generalization of our methods over the state-of-the-art supervised baselines, especially in terms of human judgments[1].

1 Introduction

Abstractive conversation summarization, which aims at processing, organizing, and distilling human interaction activities into natural, concise, and informative text (Murray et al., 2006; Wang and Cardie, 2013), is one of the most challenging and interesting tasks in text summarization. Growing attention has been paid to neural abstractive conversation summarization through a variety of designs including transferring document summarization models (Gliwa et al., 2019; Yu et al., 2021; Jia et al., 2022), utilizing conversational structures (Chen and Yang, 2020; Feng et al., 2020b; Zhu et al., 2020a; Chen and Yang, 2021b; Liu et al., 2019b; Lin et al., 2022; Zhang et al., 2022; Liu et al., 2021), introducing conversational data augmentation (Chen and Yang, 2021a), incorporating controllable signals (Narayan et al., 2021b; Wu et al., 2021) and pre-training conversation models (Zhong et al., 2021). Most of them are trained with supervised learning, which maximizes the log probability of human written summaries. While they have gained impressive performances, there are still huge gaps in generating high-quality summaries as determined by humans such as coherence or faithfulness(Chen and Yang, 2021b), largely due to a misalignment between the fine-tuning objective (*maximizing the likelihood of single human-written summary*) and the actual needs (*generating more human-favored summaries*) (Ziegler et al., 2019).

To train the summarization models on objectives that can more closely capture the behaviors humans care about, Reinforcement Learning (RL) has been used to directly optimize the rewards learned and constructed from human feedback (Ziegler et al., 2019; Stiennon et al., 2020; Böhm et al., 2019; Ye and Simpson, 2021). Different kinds of feedback have been explored to construct the reward functions such as human ratings over CNN/DM summaries (Böhm et al., 2019), overall preferences among pairs of summaries (Ziegler et al., 2019), and the similarity-redundancy matrix (Böhm et al., 2019). While achieving promising performances, they are mainly designed for document summarization with a single reward function learned from overall assessments on summaries(Böhm et al., 2019; Ziegler et al., 2019). As a result, they might not be directly adapted to dialogue summarization because of the intrinsic differences between documents and conversations. Compared to documents, conversations are generally less structured and more complex (Chen and Yang, 2020). There

[1]The data and codes are available at https://github.com/SALT-NLP/Human_in_the_Loop_Conversation_Summarization

Findings of the Association for Computational Linguistics: ACL 2023, pages 9176–9190
July 9-14, 2023 ©2023 Association for Computational Linguistics

are diverse interactions between multiple speakers and complex structures such as interruptions, discourse relations, and speaker roles in dialogues (Chen and Yang, 2020). Therefore, more subtle levels of human feedback with the consideration of *conversation structural information* is needed to provide more comprehensive, consistent, and generalizable rewards, which may lead to better performances for dialogue summarization.

To fill in this gap, we introduce Human-In-The-Loop (HITL) abstractive dialogue summarization with different levels of human feedback to leverage various conversation structures. Specifically, we incorporate two levels of human feedback: (1) **Local** Feedback, which consists of highlighted words or phrases in dialogues to capture the salient structural information, including but not limited to *speaker's intents*, *identifiable events/topics*, and *discourse relations* (e.g., causal relationships and *important emotions*), and (2) **Global** Feedback, which includes dimensions like *Coherence*, *Accuracy*, *Coverage*, *Concise* and the *Overall Quality*, to provide more comprehensive human preferences on the given summary. We hire and train human annotators to provide the introduced two levels of human feedback on 1,000 randomly sampled conversations from the DialogSum dataset (Chen et al., 2019). With the collected human feedback, we construct the **local reward** (r_l) based on the similarities between the generated summaries and the annotated highlights and learn the **global reward** (r_g) models via supervised learning which predict the human preferences. Finally, we train the summarization policy via RL to maximize the rewards predicted by r_l and r_g. Specifically, the policy generates a token of text at each time step and is updated using the PPO algorithm (Ziegler et al., 2019) based on the reward given to the entire generated summary. We conducted extensive experiments and ablation studies in different settings on the recent conversation summarization dataset, DialogSum (Chen et al., 2019) and SAMSum (Gliwa et al., 2019), to demonstrate the superiority of our methods compared to the state-of-the-art supervised learning baselines, especially in terms of human judgments and generalization abilities.

To summarize, our contributions are: (1) we introduced and collected the local and global feedback tailored for abstractive conversation summarization; (2) we designed the HITL to learn better conversation summarization policies via reinforcement learning where different levels of human feedback are directly optimized; (3) we performed extensive experiments to study the effectiveness and generation abilities of our HITL methods on DialogSum and SAMSum datasets.

2 Related Work

2.1 Abstractive Dialogue Summarization

Neural abstractive dialogue summarization has received intensive attention recently with the introduction of large-scale datasets (Gliwa et al., 2019; Chen et al., 2019; Tuggener et al., 2021). Besides directly transferring documents summarization methods to conversations (Gliwa et al., 2019), models tailored for conversation have been proposed to achieve better performances (Zhao et al., 2019; Zhu et al., 2020b; Feng et al., 2021), which make use of the rich structured information in conversations such as dialogue acts (Goo and Chen, 2018), key point/entity sequences (Liu et al., 2019a; Narayan et al., 2021a), topic segments (Liu et al., 2019c; Li et al., 2019), stage developments (Chen and Yang, 2020), discourse relations (Chen and Yang, 2021b; Feng et al., 2020a), action mentions (Zhang et al., 2022), and coreferences (Liu et al., 2021). Recent work has also explored learning in a data-efficient way through data augmentation and semi-supervised learning (Chen and Yang, 2021a), generating more controllable summaries (Wu et al., 2021; Narayan et al., 2021b). Moreover, external information such as commonsense knowledge is incorporated to help understand the global conversation context (Feng et al., 2020b). Zhong et al. (2021) pre-trained a language model on conversational data to help the summarization as well.

Most of the current dialogue summarization systems are still trained to maximize the likelihood of human-written text and have led to significant performances, but there is still a huge gap in generating high-quality summaries as determined by humans such as coherence, faithfulness, conciseness, and concreteness (Chen and Yang, 2020). This is mainly due to the misalignment between the training objective and model evaluation. For example, models never plan and look ahead for overall summarization goals. To fill in this gap, we directly learn the summarization policy that maximizes the rewards constructed from human feedback via Reinforcement Learning to generate more human-favored summaries.

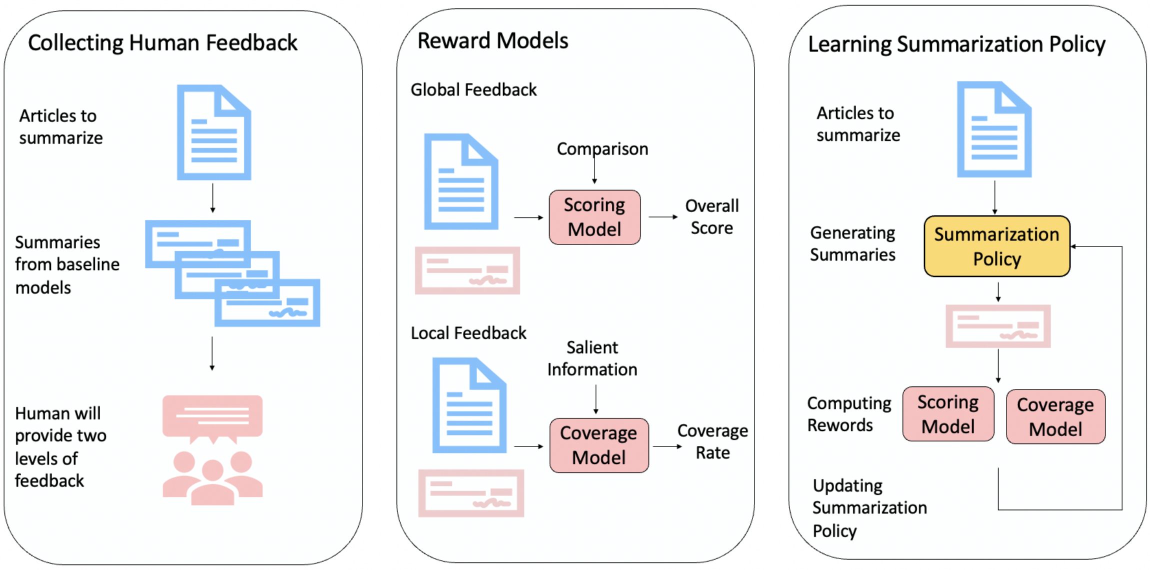


Figure 1: Overall process of our human-in-the-loop conversation summarization system including collecting human feedback, learning and designing reward models based on feedback, and learning the summarization policy.

2.2 Learning with Human Feedback

Recent research has started to explore incorporating human feedback into the training process to achieve human-preferred systems in different tasks such as dialogue generation (Jaques et al., 2019; Yi et al., 2019; Hancock et al., 2019), story generation (Zhou and Xu, 2020), document summarization (Ziegler et al., 2019; Stiennon et al., 2020; Böhm et al., 2019) and etc. Our work is most related to previous work which utilizes human feedback to train document summarization models with Reinforcement Learning (RL) (Ziegler et al., 2019; Stiennon et al., 2020; Böhm et al., 2019; Ye and Simpson, 2021), where human ratings/comparisons over summaries are usually used to learn the reward models to serve as the value networks in RL. Despite the effectiveness, it is challenging to directly apply them to conversation summarization, largely due to the complex structures in conversations, which requires subtle reward design.

Inspired by these prior work, we introduce two levels of human feedback to guide the dialogue summarization model to generate more human-favored summaries instead of only collecting pairwise-comparing binary global rating annotations, including the (1) **Local** Feedback which highlights the important conversation structures to summarize and the (2) **Global** Feedback which consists of different fine-grained dimensions to provide more comprehensive judgments. Our work is also related to using RL to optimize automatic metrics for summarization, such as ROUGE (Ranzato et al., 2015; Wu and Hu, 2018; Gao et al., 2019; Parnell et al., 2021), while we are directly optimizing human preferences with RL.

3 Methods

In this section, we introduce our **H**uman-**i**n-**t**he-**L**oop conversation summarization (**HITL**) pipeline (in Figure 1) where we incorporate two levels of human feedback, the **local** and **global** feedback, into the learning process. Inspired by Stiennon et al. (2020), our pipeline for abstractive conversation summarization includes 3 stages: (1) Collecting two levels of human feedback from conversation-summary pairs where summaries are generated with baseline models; (2) Learning and designing reward models from two levels of human feedback; (3) Learning the summarization policy which could generate higher-quality summaries as judged by humans against the reward model.

3.1 Datasets

We utilize DialogSum (Chen et al., 2019), a recent large-scale dialogue summarization dataset emphasizing real-life daily conversations, to study human-in-the-loop conversation summarization. We selected DialogSum because the summaries in DialogSum are less extractive with more novel n-grams. The summaries are more compressed compared to the other conversation summariza-

Dataset	# Turns	# Words	# Words in Sum
Sampled	9.6	127.6	22.9
DialogSum	9.8	131.0	23.6

Table 1: Data statistics of sampled 1000 dialogues and DialogSum including the average number of turns and words in conversations and the average number of words in ground truth summaries.

tion datasets (Chen et al., 2019)[2], which makes the datasets more challenging and requires human knowledge to generate better summaries.

3.2 Collecting Human Feedback

Here we describe the process of getting the desired global and local human feedback.

3.2.1 Annotation Setup

Sampling Dialogues From this DialogSum dataset, we randomly sample 1,000 dialogues from 13,360 dialogues to collect our designed two levels of human feedback. As the data statistics shown in Table 1, the distribution of our sampled examples is close to that of the original DialogSum dataset.

Baseline Summaries We generate a set of baseline summaries with different models for the global feedback annotation. Specifically, for every dialogue, we generate 4 summaries with 4 different summarization systems: (1) BART-large fine-tuned on SAMSum and XSUM [3] with a 30.4/11.5/24.8 ROUGE score on DialogSum, (2) DistilBART fine-tuned on CNN/Daily Mail and SaumSUM [4] with a 33.8/13.6/27.8 ROUGE score , (3) BART-large fine-tuned on SAMSum [5] and with a 33.0/13.5/27.0 ROUGE score (4) BART-large-xsum [6] fine-tuned on SAMSum [7] with a 26.6/10.2/21.4 ROUGE score. These different summaries are then compared by humans to provide global feedback.

Hiring and Training Annotators We hire two annotators through Upwork[8] and provide them with extensive training for the task. During multiple training sessions, we explain how to highlight salient information and compare summaries using our interfaces. We go through selected example dialogues and discuss with them to resolve inconsistencies and disagreements. To further reaffirm the training, we also perform test runs on the sampled dialogues. From these test cases, we make sure that they annotate the data properly and achieve good agreements. We pay the annotators $25 per hour. We get 41.67 hours of work for the first member and 39.83 hours for the second member [9].

3.2.2 Local Feedback

For the local feedback, we ask annotators to highlight the salient information in the provided dialogues. The highlighted information needs to be helpful in generating a summary. The information can be phrases, sentences, or a couple of words in the given dialogue. Specifically, we ask annotators to look for some important aspects including (1) *speaker's intents*, (2) *identifiable events/topics*, (3) *discourse relations* such as causal relationships and (4) important *emotions* in the conversation. For every conversation, we ask the annotator to annotate 3 to 8 highlights. After 3 rounds of training sessions, we examine the quality by asking them to annotate the same set of 50 dialogues and computing the agreement scores between the two annotators (0.865 BERT-score between their annotated spans) [10]. We also make sure the highlights match the important information in ground truth summaries (0.792 BERT-score between annotated spans and corresponding summaries) [11]. Annotators then annotate the remaining dialogues by themselves independently. After annotation, we collect 6.1 spans for every dialogue with 59.5 words on average.

3.2.3 Global Feedback

After highlighting the salient information, we provide the annotators with 3 pairs of summaries sampled from the set of baseline summaries. We then ask them to make comparisons in terms of *Coherence, Accuracy, Coverage, Conciseness*, and *Overall Quality*. For every comparison between summary A and summary B, the annotators need to grade on a scale of 5 points: summary A is mostly

[2]The data statistics are shown in Table 7 in the Appendix.

[3]https://huggingface.co/Salesforce/bart-large-xsum-samsum

[4]https://huggingface.co/philschmid/distilbart-cnn-12-6-samsum

[5]https://huggingface.co/linydub/bart-large-samsum

[6]https://huggingface.co/facebook/bart-large-xsum

[7]https://huggingface.co/knkarthick/meeting-summary-samsum

[8]https://www.upwork.com

[9]The interface is shown in the Appendix

[10]The BERTScore for randomly sampled pairs of spans is 0.573.

[11]The BERTscore for randomly sampled pairs of utterance and summary is 0.469.

better, summary A is partially better, equal, summary B is partially better, the summary B is mostly better. We provide detailed guidelines to the annotators about those different dimensions[12]. After 3 rounds of training sessions, we show the annotators 50 dialogues with 150 pairs of summaries and ask them to make comparisons, resulting in 150 comparisons. We then calculate the Fleiss Kappa scores to measure the agreements among different annotators. In the end, we obtain an average score of 0.342 for Coverage, 0.381 for Coherence, 0.376 for Conciseness, 0.373 for Accuracy, and 0.369 for Overall Quality, indicating moderate agreement (Landis and Koch, 1977). Annotators then annotate the remaining dialogues by themselves independently. In total, we collect 3000 pairs of comparisons for every dimension.

3.3 Method

This section focuses on how to incorporate the annotated feedback into the training process to assist the summarization systems in generating more human-favored summaries.

3.3.1 Rewards Modeling

We first describe how to train the reward models and compute the rewards for any given conversation-summary pairs.

Local Rewards Our goal is to encourage the summarization systems to generate summaries that cover the important information mentioned in the dialogues while avoiding redundant information. Thus here we propose to model the local rewards based on these highlights from annotators. For a given conversation C with a set of human-annotated salient spans $M = M_{1:m}$ (e.g., phrases/sentences/words in the dialogues), suppose the model would generate a summary s. We view the list of highlights M annotated by humans as information needed by the summaries, and the other sentences without highlights as possible redundant information set $N = N_{1:n} = C - M$. We then calculate the local coverage rewards $r_l(C, s, M)$ by calculating the cosine distances between the embeddings of the summary and the information in the dialogues:

$$r_l(C, s, M) = \sum_{i}^{m} cos(s, M_i) - \sum_{j}^{n} cos(s, N_j) \quad (1)$$

[12]The detailed guidelines are shown in the Appendix.

Here we embed the summaries and the dialogue information utilizing sentence-transformers (all-mpnet-base-v2) [13] (Reimers and Gurevych, 2019).

Global Rewards Generating high-quality summaries with better human preferences is essential for building better summarization systems. To this end, we design the global rewards by learning human preferences from their annotations. For a given set of annotated conversations $C = \{C_1, ..., C_n\}$ with baseline summaries $S = \{(s_1^1, s_2^1, s_3^1, s_4^1), ..., (s_1^n, s_2^n, s_2^n, s_3^n)\}$ with different dimensions of global human feedback, we first learn a set of reward models $r_{g_j}(C, s; \theta_e, \theta_j)$ to measures the quality or impact of the dimension j on summary s for a conversation C. Here, θ_e are the parameters to encode the conversation and summaries; θ_j stands for the parameters of linear heads for the dimension j, which outputs a scalar on top of θ_e. Specifically, we initialize θ_e with a BART-large model fine-tuned on the DialogSum dataset, and randomly initialize θ_j for every dimension in the global feedback. During training, we train the model to predict which summary in a summary pair $\{s_n^i, s_m^i\}$ of conversation C_i is preferred by humans, by optimizing the following objective:

$$\mathcal{L} = -\mathbb{E}_{(C_i, s_n^i, s_m^i) \sim (C,S)} \Sigma_j [\log(\sigma(r_{g_j}(C_i, s_n^i; \theta_e, \theta_j) - r_{g_j}(C_i, s_m^i; \theta_e, \theta_j)))]$$

where s_n^i is the summary preferred by humans.

Implementations are shown in Section 4.2. We select the hyper-parameter based on the loss on the validation set (8:2 split), and further evaluate the learned reward models in Section 4.4.

We then combine different dimensions to provide the global rewards $r_g(C, s)$:

$$r_g(C, s) = \Sigma_j r_{g_j} \quad (2)$$

3.3.2 HITL Summarization Policy Learning

Here we train a summarization policy with human feedback for generating high-quality outputs as judged by humans. We utilize reinforcement learning to learn the summarization policy π_ϕ^{RL}. Specifically, we initialize the policy with a supervised learning BART-large baseline model π^{B} fine-tuned on DialogSum. We use the PPO algorithm (Schulman et al., 2017) to maximize the rewards from the above local and global reward models r_l and r_g,

[13]https://github.com/UKPLab/sentence-transformers

Methods	# Training Data	Rewards	ROUGE-1	ROUGE-2	ROUGE-L
BART-large	Full	-	47.28	21.18	44.83
HITL-synthesis	Full	r_g	46.87	21.03	45.12
HITL-synthesis	Full	r_l	47.27	22.18	45.15
HITL-synthesis	Full	$r_g + r_l$	47.46	22.13	45.24
HITL-synthesis	1000	r_g	46.25	20.79	44.37
HITL-synthesis	1000	r_l	46.18	21.12	45.13
HITL-synthesis	1000	$r_g + r_l$	46.38	21.26	45.08
HITL†	1000	r_g	47.54	23.05	45.38
HITL†	1000	r_l	47.88	23.17	45.87
HITL†	1000	$r_g + r_l$	**48.29**	**23.65**	**46.23**

Table 2: ROUGE-1, ROUGE-2 and ROUGE-L scores for different models on the DialogSum Corpus test set. † means our model. We performed Pitman's permutation test (Dror et al., 2018) and found that *HITL* significantly outperformed the supervised baseline *BART-large* ($p < 0.05$). The results are averaged over three random runs.

Methods	Human Preferred %
BART-large	18%
HITL-($r_g + r_l$) †	**82%**

Table 3: Human preferences when comparing summaries generated by supervised baseline model (BART-large) and our best HITL model ($r_g + r_l$). † means our method.

where each time step is a BPE token [14]. The full reward $R(C, s, M)$ is:

$$R(C,s,M) = w_l r_l(C,s,M) + w_g r_g(C,s) - \beta \log \left[\frac{\pi_\phi^{\mathrm{RL}}(s|C)}{\pi^{\mathrm{B}}(s|C)} \right] \quad (3)$$

We introduce a KL divergence term between the HITL policy and the supervised baseline model (Jaques et al., 2019). This term could prevent the learned policy generating outputs that are too different from the supervised models and encourage the learned policy to explore instead of collapsing to a single model (Stiennon et al., 2020). w_l, w_g and β are weights to balance different sub-rewards.

Following Stiennon et al. (2020), we use a Transformer with separate parameters from the policy for the PPO value function. And we initialize the value function to the parameters of the reward model. In our experiments, the reward model, policy and value function are the same size.

[14]The reward model would give the rewards after the entire summary generated. Each episode terminates when the policy outputs the EOS token, and the discount factor $\gamma = 1$.

Metric	Agree with Human %
ROUGE	55.3%
Coherence	62.4%
Accuracy	56.8%
Coverage	63.6%
Concise	59.5%
Over Quality	65.5%
r_g	**69.8%**

Table 4: Agreement with human preferences of different reward models including the ROUGE score, coherence reward model, accuracy reward model, coverage reward model, concise reward model, over quality reward model and our global reward model r_g.

4 Experiments

4.1 Baselines

We compare our models with several baselines:

- **BART-large** (Lewis et al., 2020): We utilized BART-large as our backbone model as well as the supervised learning baseline. Utterances are separated by a special token.

- **HITL-synthesis**: We use heuristics to approximate the local and global feedback, via which we then learn synthesized reward models and the HITL summarization policy. Specifically, for the local feedback, we utilize a greedy algorithm (Nallapati et al., 2016; Zhang et al., 2022) to obtain the synthesis highlights based on ground truth. For the global feedback, we

Methods	Training Data	Transferred Parts	ROUGE-1	ROUGE-2	ROUGE-L
BART-large	DialogSum	Whole Model	31.74	5.93	29.79
HITL-($r_g + r_l$)†	DialogSum	Whole Model	**33.58**	**7.84**	**32.63**
BART-large	SAMSum	Whole Model	53.12	27.95	49.15
HITL-(r_g) † †	SAMSum	Global Reawrds	**53.76**	**28.04**	**50.56**

Table 5: ROUGE-1, ROUGE-2 and ROUGE-L scores on the *SAMSum test data*. † means that we directly apply our HITL models trained on DialogSum to SAMSum. † † means that we re-train the policy on SAMSum corpus with the global reward learned from DialogSum annotations. The results are averaged over three random runs.

Quality	R1	R2	RL
Synthesis	46.38	21.26	45.08
Noisy	46.32	21.38	44.76
Clean	**47.58**	**22.58**	**45.56**

Table 6: ROUGE scores on the DialogSum test data where the HITL-($r_g + r_l$) policy is learned with 400 annotations with different qualities. The results are averaged over three random runs.

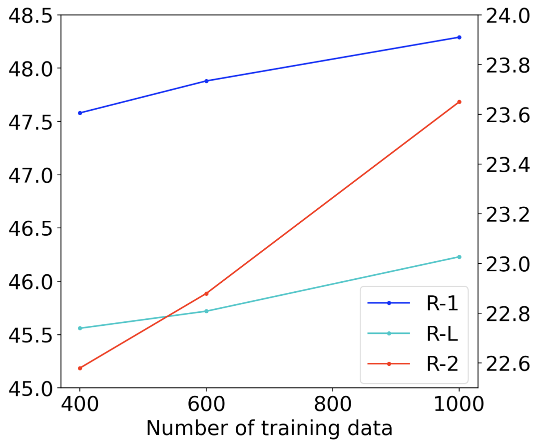


Figure 2: ROUGE-1, ROUGE-2 and ROUGE-L scores on the DialogSum test data where the HITL policy is learned with different number of annotations (400, 600, 1000). The left y axis means ROUGE-1 and ROUGE-L, the right y axis means ROUGE-2.

utilize the randomly sampled utterances as negative summaries compared to the ground truth summary.

4.2 Implementation Details

For the supervised baseline, we initialize the model with BART-large and fine-tune it on the full DialogSum for 10 epochs with a 3e-5 learning rate and 120 warm-up steps. We use a batch size of 8. For the global reward models, we set the hidden size of the linear head 256. We use a batch size of 8 and train the reward model for 2 epochs with a 3e-5 learning rate. For PPO, we initialize our policies with the supervised baseline and our value functions with the reward models. We set $\gamma = 1$ and $\lambda = 0.95$ for the advantage estimation (Schulman et al., 2015), do 4 epochs of optimization with a batch size of 8 and run for 5,000 episodes. We set $w_l = 1$, $w_g = 1.5$ and $\beta = 0.05$ based on grid search among $\{0.05, 0.5, 1, 1.5, 2, 2.5\}$ for the full reward R. All experiments were performed on 8 NVIDIA V100 (32GB memory).

4.3 Automatic Evaluation

We first evaluated all the models with the widely used automatic metric ROUGE(Lin and Och, 2004) and reported ROUGE-1, ROUGE-2, and ROUGE-L in Table 2. We found that the performances were not better for synthesis data when there were less training data. When there was plenty of synthesis feedback, (*HITL-synthesis with Full data*) can help improve over the supervised baseline, where the local reward was more important compared to the global reward. After incorporating ground-truth human feedback, our *HITL-($r_g + r_l$)* model with both global and local rewards achieved the best performances even with less training data compared to synthesis baselines. The local rewards consistently brought in more performance boost because the conversation structural information in local rewards can help the systems more directly capture the important factors in the conversation. This indicates the effectiveness of our HITL framework for conversation summarization, as the human judgements were directly guiding the learning process.

4.4 Human Evaluation

Following Böhm et al. (2019) and Stiennon et al. (2020), we randomly sampled 200 conversations from the DialogSum test set and asked annotators from Amazon Mechanical Turk to select the preferred summary from pairs of summaries generated by *BART-large* and *HITL-*$(r_g + r_l)$. Turkers were asked to judge coherence, accuracy, coverage, and conciseness [15]. To increase the annotation quality, we required Turkers to have a 98% approval rate and at least 10,000 approved tasks for their previous work. Each conversation was rated by three workers, and we used majority voting to decide the preferred summaries. The pay rate was 0.5$ per hit. We measured the agreement by computing the Intra-Class Correlation (ICC) was 0.693, showing moderate agreement (Koo and Li, 2016).

Main Results From Table 3, we observed that summaries from our introduced (*HITL-*$(r_g + r_l)$) are much more preferred (favored in **82%** cases) by human compared to supervised baseline (*BART-large*). These significant improvements came from comparably *small amount of annotations* (1000 dialogues). These indicated that the systems (*HITL-*$(r_g + r_l)$) that directly learn from a small amount of global and local human feedback could generate higher-quality summaries with better human preferences compared to supervised baselines.

Evaluating the Global Reward Models Based on human preferences, we further examined the global reward model and compared it with its subdimensions as well as the ROUGE metric. Basically, we assume that the reward model agrees with human preferences when the model is assigning higher scores to these human-preferred summaries. As shown in Table 4, reward models learned from humans generally agree well with humans, where our global reward r_g receives the highest agreement rate. This showed the high quality and effectiveness of our global feedback collection as well as the global reward models. As a result, our *HITL-*$(r_g + r_l)$ model achieves better performances compared to baselines. The results also showed the potential of our global reward models to be used to better automatically evaluate the summaries (Fabbri et al., 2020).

4.5 Generalization

We then evaluated the generalization abilities of our *HITL* summarization system and our learned global reward model r_g. We transferred the knowledge learned on DialogSum to another corpus, SAMSum (Gliwa et al., 2019) which summarizes messenger-like conversations about daily topics, such as arranging meetings and discussing events. The good generalization shown below also lower the amortized cost (Rajani et al., 2021) of our methods.

Generalization of HITL models We first directly applied the whole *HITL-*$(r_g + r_l)$ models trained on DialogSum to the SAMSum corpus. The results were visualized in Table 5. The zero-shot evaluations on SAMSum got lower ROUGE scores compared to the models trained on SAMSum data, while our best model, *HITL-*$(r_g + r_l)$, achieved better performances compared to the supervised baseline model (*BART-large*). This showed that our policy empowered with human feedback can better generalize to other domains compared to supervised learning models because our policy was learned from rewards that explicitly indicated human preferences. Such rewards are more general to different domains compared to supervised learning objectives which are specific to one dataset.

Generalization of the Global Reward Model We then re-trained the *HITL-*(r_g) policy on SAMSum corpus while we directly utilized the global reward model $r_g(C, s)$ learned from human feedback on DialogSum data as the reward functions. We reported the results in Table 5 and observed that the *HITL-*(r_g) outperformed the supervised BART-large model on SAMSum in terms of ROUGE scores. This showed that our global reward models r_g can be directly applied to other conversation summarization datasets to provide reinforcement learning rewards and boost performance because the global rewards learned from human feedback are implying the qualities of summaries in general rather than being limited to one specific domain.

4.6 Ablation Study

Here we performed two ablation studies to further study the impact of the quality and the quantity of human feedback in our HITL pipeline.

Perturbing the Qualities of Annotations We compared *HITL-*$(r_g + r_l)$ policy trained with annotations on the same 400 dialogues of three levels of qualities: (1) *Synthesis Annotations* as described in Section 4.1, (2) *Noisy* which was the annotation from annotators without extensive training sessions, (3) *Clean* which was the annotation after the training sessions. We visualized the comparisons in

[15]The guidelines are the same as the guidelines for global feedback annotations described in the Appendix.

Table 6, and found that the performances were better with higher quality annotations. This suggests that the quality of human feedback matters.

Increasing the Annotations We then varied the number of annotations from 400 to 1000 in our *HITL-*($r_g + r_l$) model in Table 2. The ROUGE scores were higher with more human annotations because of better reward learning and policy learning with more training data. This implies the importance of enough human feedback to learn and design better rewards.

5 Conclusion

In this work, we introduced two levels of conversation human feedback into the abstractive conversation summarization to generate human-preferred summaries. Specifically, we first collected local and global human feedback to design the corresponding reward functions. We then learned the summarization policies via reinforcement learning to optimize the designed rewards. Extensive experiments in different settings and ablation studies on DialogSum and SAMSum corpus via both automatic and human evaluations demonstrated the effectiveness and generalization of our introduced HITL pipeline. For future work, we would like to explore incorporating human feedback in natural languages which are more general and explicit to indicate how to summarize conversations to improve the abstractive conversation summarization.

6 Limitation

In this work, we collect extensive and comprehensive human feedback with high qualities to facilitate our human-in-the-loop conversation summarization framework. While the learned rewards and models are showing good generalization abilities, further attention is still needed to deeply understand what types of feedback or what amount of feedback is necessary. Our current work only considers human feedback collected using the required forms (i.e., rankings and highlighting). We encourage future work to explore how to incorporate human preferences with more open-ended feedback such as through natural languages. Furthermore, we mainly focus on conversation summarization with human feedback in this work, and other types of summarization tasks (e.g., multi-document summarization, email to-do summarization, meeting summarization and etc.) could be further explored to incorporate human knowledge.

Acknowledgements

We thank members of the SALT Lab and the anonymous reviewers for their helpful feedback. This work was supported in part by an Amazon Faculty Research Award and an NSF grant IIS-2247357.

References

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references.

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 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic. 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.

Jiyu Chen, Karin Verspoor, and Zenan Zhai. 2019. A bag-of-concepts model improves relation extraction in a narrow knowledge domain with limited data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 43–52, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020a. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020b. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*.

Yang Gao, Christian M. Meyer, Mohsen Mesgar, and Iryna Gurevych. 2019. Reward learning for efficient reinforcement learning in extractive document summarisation.

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.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *2018 IEEE Spoken Language Technology Workshop (SLT)*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazaré, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot!

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog.

Qi Jia, Yizhu Liu, Haifeng Tang, and Kenny Q. Zhu. 2022. Post-training dialogue summarization using pseudo-paraphrasing.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD19, page 1957–1965, New York, NY, USA. Association for Computing Machinery.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, et al. 2019b. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 24–31.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019c. Topic-aware pointer-generator networks for summarizing spoken conversations. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Zhengyuan Liu, Ke Shi, and Nancy F. Chen. 2021. Coreference-aware dialogue summarization.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 367–374, New York City, USA. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021a. Planning with entity chains for abstractive summarization.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, Vitaly Nikolaev, and Ryan McDonald. 2021b. Planning with learned entity prompts for abstractive summarization.

Jacob Parnell, Inigo Jauregi Unanue, and Massimo Piccardi. 2021. Rewardsofsum: Exploring reinforcement learning rewards for summarisation.

Vineet Rajani, Marco Gaboardi, Deepak Garg, and Jan Hoffmann. 2021. A unifying type-theory for higher-order (amortized) cost analysis. *Proc. ACM Program. Lang.*, 5(POPL).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback.

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning.

Yuxuan Ye and Edwin Simpson. 2021. A proposal: Interactively learning to summarise timelines by reinforcement learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 25–31, Online. Association for Computational Linguistics.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Z. Hakkani-Tür. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *INLG*.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904, Online. Association for Computational Linguistics.

Kexun Zhang, Jiaao Chen, and Diyi Yang. 2022. Focus on the action: Learning to highlight and summarize jointly for email to-do items summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4095–4106, Dublin, Ireland. Association for Computational Linguistics.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, Min Yang, and Deng Cai. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference*, WWW '19, page 3455–3461, New York, NY, USA. Association for Computing Machinery.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Dialoglm: Pre-trained model for long dialogue understanding and summarization.

Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models.

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

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020b. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences.

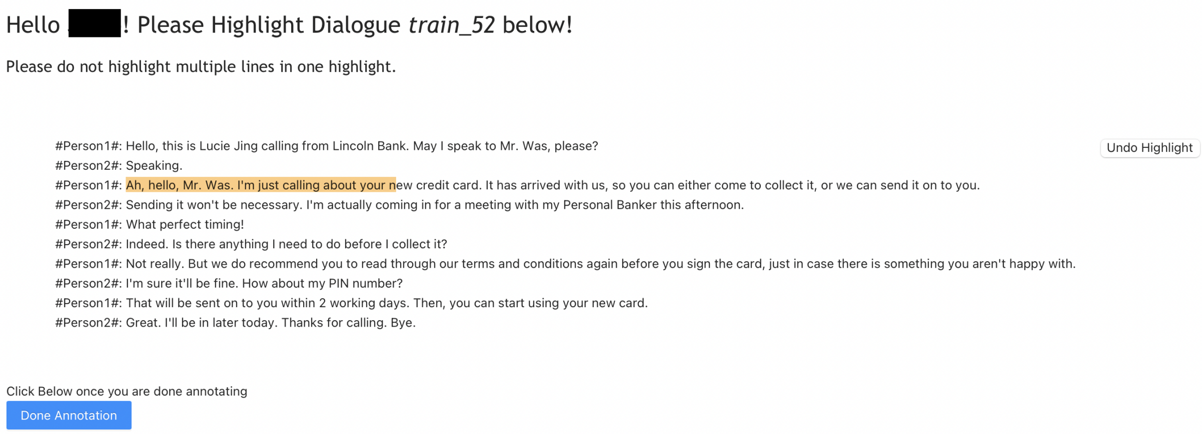
A Data Statistics for DialogueSUM and SAMSum

B The Annotation Interface

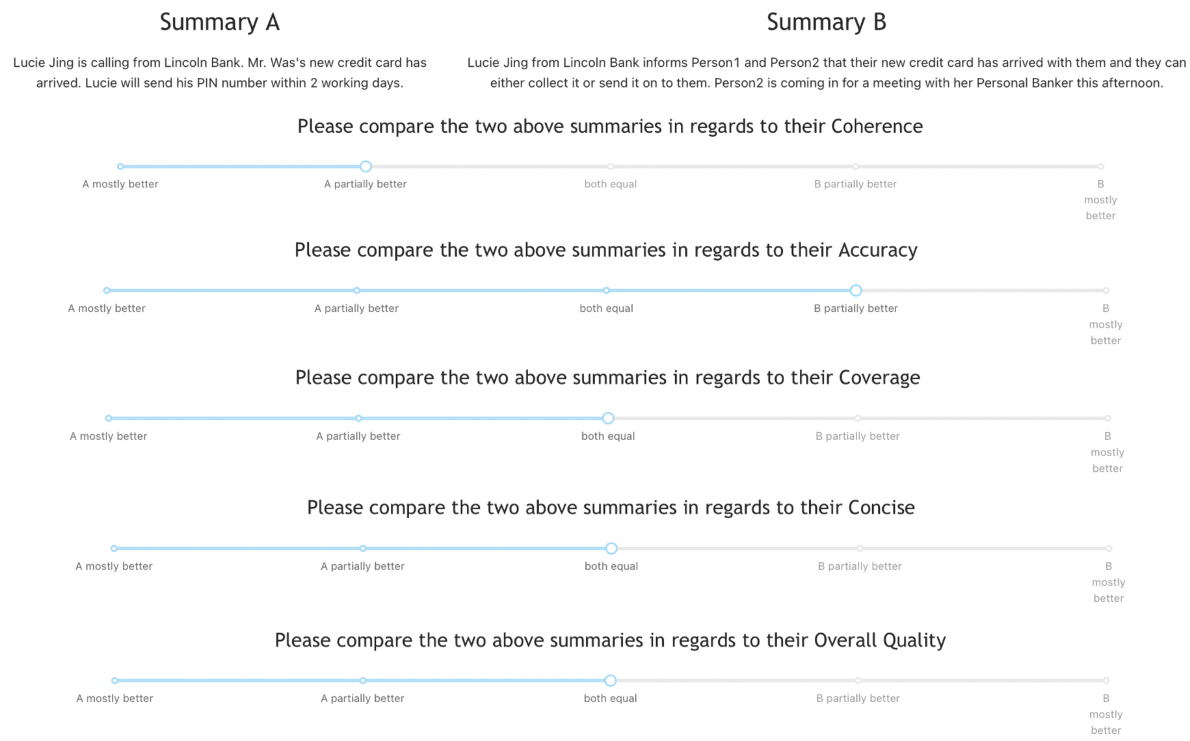
Since we hired and trained our own set of annotators, rather than using a crowd sourcing website such as Amazon Mechanical Turk, we built

Dataset	# of Dialogues	Avg # of turns	Avg # of Words	Avg Compression Rate
DialogSUM	13,406	9.5	131.0	0.18
SAMSum	16,369	11.1	94.3	0.30

Table 7: Data Statistics of DialogSUM and SAMSum



(a)



(b)

Figure 3: The designed websits to collect data: (a) Highlighting key information in a given conversation. (b) Comparing two given summaries in terms of given aspects.

our own website to allow for a standardized, customized user interface for all annotators. The website contains the information for highlighting, summary comparisons as well as detailed instructions. From here we collect local and global guidance. For local guidance, we display one of the dialogues on the website. We ask the user to highlight salient information and then press next. Afterward, we display 3 pairs of summaries and ask the user to compare the pairs of summaries in 5 different dimensions. Screenshots from the website are shown in Figure 3. Data collected from the website can be easily ported into a central database containing all of our human data.

C Global Feedback Guidelines

We provide the annotators with 3 pairs of summaries sampled from the set of baseline summaries, and ask them to make comparisons in terms of *Coherence, Accuracy, Coverage, Conciseness*, and *Overall Quality*. For every comparison between summary A and summary B, the annotators need to grade upon a scale of 5 points: summary A mostly better, summary A partially better, equal, summary B partially better, summary B mostly better. We provide detailed guidelines to the annotators about those different dimensions:

- **Coherence**: Summary is easy to understand and free of English errors. For comparing summaries against each other in Coherence, we ask the annotators to compare the number and severity of grammatical, syntax, and spelling errors of each summary against each other.

- **Accuracy**: Information that stated in the summary is accurate and does not incorrect information. Summary is not misleading and has too much errors. For comparing summaries against each other in Accuracy, we ask the annotators discover the amount and severity of inaccurate statements that occur in the summaries against each other.

- **Coverage**: Mentions main information of the conversations. It conveys the most salient information from the dialogue. For comparing summaries against each other in Coverage, we ask the annotators to look at the number of events in each summary. Also taking into the factor of importance of events, we ask the annotator to compare the number of events against the pair of summaries.

- **Conciseness**: Summary is short and to the point. It does not have too much unimportant information that is not included in the salient information. For comparing summaries against each other in Conciseness, we ask the annotators to mainly look at the length of the summaries. Then we check if any information doesn't fit, and penalize as such.

- **Overall Quality**: We ask the annotator to use all of the above information and other related context to give an overall rating. Even though we asked the annotator to consider all the information, we asked the annotator to factor coverage and accuracy more into their decision for Overall Quality. This is because it is of at most importance for a dialogue summary to accurately summarize the salient information of the dialogue.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.