Summary Grounded Conversation Generation

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Sachindra Joshi, David Konopnicki
IBM Research AI
chulaka.gunasekara@ibm.com, {guyf, benjams}@il.ibm.com
{jsachind@in, davidko@il*}.ibm.com

Abstract

Many conversation datasets have been constructed in the recent years using crowd-sourcing. However, the data collection process can be time consuming and presents many challenges to ensure data quality. Since language generation has improved immensely in recent years with the advancement of pre-trained language models, we investigate how such models can be utilized to generate entire conversations, given only a summary of a conversation as the input. We explore three approaches to generate summary grounded conversations, and evaluate the generated conversations using automatic measures and human judgements. We also show that the accuracy of conversation summarization can be improved by augmenting a conversation summarization dataset with generated conversations.

1 Introduction

Automatic conversation systems require large quantities of data to learn task specific language patterns and underlying conversation policies. Such data either come from human-to-human conversation logs (Lowe et al., 2015; Hardalov et al., 2018) or is collected in crowd-sourced environments, where two or more crowd-workers play specific roles under some guidelines (Zhang et al., 2018; Budzianowski et al., 2018). Since real human-to-human conversation logs are scarce, many datasets have been created using the latter approach. However, crowd-sourced conversation data collection is time consuming, costly and presents multiple challenges to ensure data quality (Kang et al., 2018).

Conversation summarization is an emerging research area that has been ill-studied due to the lack of large-scale datasets. Most existing public datasets in this domain are small, for example, AMI meeting corpus (McCowan et al., 2005) contains 137 summary transcripts. CRD3 (Rameshkumar and Bailey, 2020) is a spoken conversation dataset that consists of 159 conversations and summaries. Samsum (Gliwa et al., 2019), the only large scale dataset for conversation summarization, contains over 16,000 open-domain conversations and summaries created artificially by humans.

Large scale pre-trained language models (PLMs) (Lewis et al., 2020; Brown et al., 2020; Raffel et al., 2020) have been used in various text generation tasks (Budzianowski and Vulić, 2019; Min et al., 2020; Cachola et al., 2020). In recent studies, PLMs are used to generate training data for natural language processing (NLP) applications. For example, Anaby-Tavor et al. (2020); Yang et al. (2020) use PLMs to create paraphrases for intent classifiers in conversation systems, and show that, when the original datasets are augmented with the generated data, performance improves. More recently Mohapatra et al. (2020) generated entire conversations grounded on instructions that are provided to crowd-workers using a modular approach, where different PLMs are trained for different roles.

Our Contributions: We investigate how PLMs can be utilized to generate entire conversations that are grounded on a given summary. We explore three approaches: (1) Supervised Learning (SL) based conversation generation *(SL-Gen)*: where, a PLM is trained to generate an entire conversation, taking the summary of a conversation as input, (2) Reinforced Learning (RL) based conversation generation *(RL-Gen)*: where, we further improve the *SL-Gen* method using the quality of the generated conversations as a reward, and (3) Controlled turn-by-turn conversation generation *(CN-Gen)*: which allows us to generate conversations turn-by-turn, constrained on the summary and a set of pre-defined control parameters. We evaluate the quality of the generated conversations by conducting automatic and human evaluation. We

*Current address: david.konopnicki@booking.com

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3748–3756

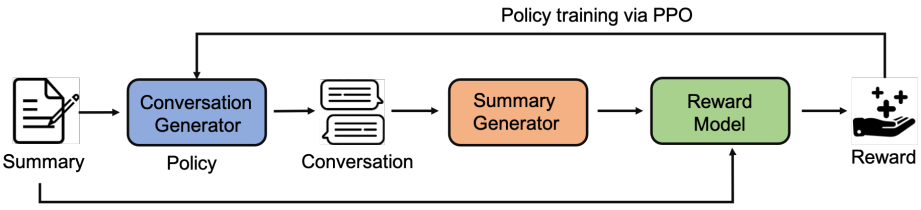

Figure 1: The RL based conversation generation framework

also show that once a conversation summarization dataset is augmented with the generated conversations, the performance of the downstream summarization task is improved.

2 Summary grounded conversation generation

In the conversation summarization task, a model takes a conversation as input, and learns to generate a summary. We study the inverse of that problem, where the input to our model is a summary, and the model generates a conversation. In this section, we propose three models for this task and the hyperparameters used in training the models are available in Section A of the appendix.

2.1 SL based generation (SL-Gen)

A seq2seq model can be trained for this task by providing a summary as the input and generating a conversation token-by-token. As PLMs have shown significant improvement over the traditional seq2seq architecture for text generation, we use a GPT-2 model and fine-tune it to generate a conversation given a summary as the input. Our input to the model follows the following format: *<bos>summary text <dialog>conversation text<eos>*. We also use different token-type-ids to indicate the summary and the conversation text. The model is trained to optimize Cross Entropy loss.

2.2 RL based generation (RL-Gen)

Many studies train text generation models with RL (Paulus et al., 2018; Li et al., 2016), where the generator network is optimized with a task specific reward. We investigate how the quality of the generated conversation can be used as a reward to improve the generation network. To this end, we train a summary generator network, which generates a summary, given a conversation. We measure the quality of the generated conversation by identifying the similarity between the summary of the generated conversation (generated, in turn, by the summary generator network) and the ground truth summary. The similarity score is used as a reward to train the conversation generation model. Our RL based generation framework is shown in Figure 1, and the critical components are described below.

Conversation Generator: A trained SL-Gen model is used as the conversation generator, which, given an summary can generate a conversation.

Summary Generator: We use a lightweight variant of BART (Lewis et al., 2019), named *Distil-BART*, which is fine-tuned on the Extreme summarization task (Narayan et al., 2018). We further fine-tune this instance on the conversation summarization data by providing the conversations as the input and training the model to output summaries.

Reward Model: Once the Summary Generator generates an output summary for the generated conversation, the reward model compares it with the ground truth summary, which was used to ground the conversation generation. As Paulus et al. (2018) we use ROUGE-2 F1-score as the reward.

Policy training: We use proximal policy optimization (Schulman et al., 2017) as the optimizer for the policy training as it prevents the generator from deviating far away from the pretrained LM (Wu et al., 2020).

2.3 Controlled conversation generation

We propose another approach, *(CN-Gen)*, for conversation generation, which grants more control over the properties of the generated conversations. Here, we generate one utterance of the conversation at a time, as opposed to the *RL-Gen*, where we generate the whole conversation at once. The properties of the generated conversations is controlled by adding several components to the input sequence to the model. The following three variables were used as the control parameters, (1) Number of remaining turns to generate in the conversation (Num turns): During the generation of a turn, we indicate the remaining number of turns in the conversation. In generating a n turn conversation, this starts with n for the first turn and reduces by 1 after the generation of each turn, (2) The speaker of the next

Summary: person0 will be late. person1 will order pasta with salmon and basil for her.

2 turn conversation:

<Person0> I'll be late
<Person1> I'll order some pasta with salmon and basil for you.

3 turn conversation

<Person0> I'll be late.
<Person1> I'll order some pasta with salmon and basil for you.
<Person0> Thanks a lot!

6 turn conversation

<Person0> Hello, I am going to be late.
<Person1> Ok
<Person1> I'll order some pasta with salmon and basil
<Person0> Ok, sounds good!
<Person0> Thank you!
<Person1> No problem

10 turn conversation

<Person0> I'll be late
<Person1> ok
<Person1> do you want me to order something for you?
<Person1> pasta?
<Person0> Yes
<Person1> with salmon?
<Person0> Yes
<Person1> Ok
<Person1> how about basil?
<Person1> Yes please!

Table 1: Multiple conversations generated by the CN-Gen approach grounded on the same summary

turn (Speaker): This indicates to the model the speaker of the next turn, and (3) The length of the next turn (Turn length): We define, 3 categories of lengths: Short ($\leq$ 3 tokens), Long ($>$ 10 tokens) and Medium (otherwise).

We use the following input representation to fine-tune a GPT-2 model: *<bos> summary text <context> dialog context <turns_to_go> Num turns <speaker> speaker <turn_length> turn length <turn> utterance <eos>*. Changing these parameters allows us to generate different variants of conversations which are grounded on the same summary. During training, we obtain the values for the control parameters from the ground truth conversations, and at inference we randomly select the next speaker, number of turns of the conversation to be generated (in a range of 4-15 turns), and the next turn length. In Table 1 we show conversations of different lengths that were generated by the CN-Gen approach grounded on the same summary by changing the control parameters.

A summary and a conversation from the Samsum dataset (Gliwa et al., 2019), along with the conversations generated by the three aforementioned algorithms are shown in Figure 2. More examples are provided in the Section B of the Appendix.

3 Experiments

We experiment on the Samsum (Gliwa et al., 2019) dataset, which, to the best of our knowledge, is the only public large-scale conversation summarization dataset. We pre-process the dataset by replacing the personal names (ex: John) with unique tags (ex:<person_0 >). First, we evaluate of the quality of generated conversations using automatic measures and human judgments, and then assess the performance of the generated conversations in a downstream summarization task after augmentation.

Model	Ave. Turns	Ave. Tokens/Turn
Ground truth	11.55 ± 6.48	7.10 ± 6.29
SL-Conv-Gen	10.54 ± 6.80	5.69 ± 4.40
RL-Conv-Gen	8.40 ± 4.78	5.14 ± 3.64
CN-Conv-Gen	9.70 ± 5.67	5.62 ± 4.05

Table 2: Properties of the generated conversations.

3.1 Quality of the generated conversations

We evaluate the quality of the conversations generated by the three approaches that were introduced in Section 2. In Table 2 we show the properties of generated conversations and the ground truth conversations in the test set of Samsum dataset.

Automatic Evaluation: We trained the conversation generation models on the Samsum training set and generated conversations on the test set. We compare the generated conversation with the ground truth conversations using the measures used by Sharma et al. (2017) to evaluate conversation system responses. The results shown in Table 3 suggest that CN-Gen outperform the SL-Gen and RL-Gen on all measures.

We also compare the summaries of generated conversations (generated by the Summary Generator) with the ground truth summaries, and the results are shown in Table 4. We believe that this is a semantic evaluation of the conversations, as the summaries capture the crux of the conversations. According to the results, CN-Gen outperforms the other two methods. This, along with the previous result suggest that the conversations produced by CN-Gen are the most similar to the ground truth conversations.

Human Evaluation: To evaluate the quality of generated conversations, we randomly selected 50 summaries from the Samsum test dataset and generated conversations using the three models. Three NLP experts were then asked to read the ground truth summary and rank the four conversations (3 generated and the ground truth conversation) using a [1-5] scale according to Grammaticality, Coherency, and Informativeness, with respect to the ground truth summary. Results are shown in table 5. As expected, the ground-truth conversations obtained the highest scores on all three aspects and can be considered as an upper bound for this task.

Summary: person_0 is inviting person_1 for Easter. person_1 will bring some chocolate eggs.

Ground Truth Dialog:	SL-Gen:	RL-Gen:	CN-Gen (5 turns):	CN-Gen (9 turns):
<person_0> Hi darling, do you want to come for Easter?	<person_0> Hi person_1!	<person_0> How are you doing?	<person_0> I'm inviting you over for Easter ;)	<person_0> Hey, how are you?
<person_1> Love to, I'm off on Friday	<person_1> Hi person_0!	<person_1> Good, and you?	<person_1> Yay!	<person_1> Doing good. And you?
<person_0> it's could be nice, I'll invite Louise too	<person_0> Are you still coming this Easter?	<person_0> Not bad. You?	<person_0> What can you bring?	<person_0> I'm fine
<person_1> great, I'll bring you eggs, chocolate ones of course!	<person_1> Oh, yes! I'm coming this weekend.	<person_1> Not bad.	<person_1> Ok, I'll bring some chocolate eggs, what do you think?	<person_0> Do you have any plans for Easter?
<person_0> Thanks darling.	<person_0> Ok, I'll bring some chocolate eggs and some cake	<person_0> Are you doing anything special for Easter?	<person_0> You're so sweet <3	<person_1> Not really.
	<person_1> Great!	<person_1> Hmm... I don't know.		<person_0> Would you like to come over?
		<person_0> Don't worry. I'm inviting you.		<person_1> I would love to.
		<person_1> Oh, cool.		<person_1> I'll bring some chocolate eggs.
		<person_0> What do you want to bring?		<person_0> Awesome!
		<person_1> I'll bring some chocolate eggs.		
		<person_0> Good idea!		
		<person_1> No problem.		

Figure 2: Examples of a conversations grounded on the same summary. The key terms are highlighted in colors.

Model	BLEU-4	METEOR	ROUGE-L
SL-Gen	2.81	12.06	21.53
RL-Gen	3.53	12.29	25.40
CN-Gen	**4.94**	**15.64**	**26.22**

Table 3: Evaluation of generated conversations against ground truth conversations

Model	ROUGE_1	ROUGE_2	ROUGE_L
SL-Gen	46.85	25.29	45.97
RL-Gen	52.51	31.23	51.68
CN-Gen	**53.46**	**32.52**	**52.93**

Table 4: Rouge F1 evaluation of summaries of conversations against the ground truth summaries

RL-Gen and CN-Gen obtained higher scores than SL-Gen and relatively good scores compared to the Ground Truth conversations. This corroborates the assumption that our proposed models generate high quality conversations. The Welch Two Sample t-test (Welch, 1947) shows that both RL-Gen and CN-Gen models outperform the SL-Gen model statistically significantly with $p < 0.0001$. However, there is no statistical significance between the results obtained from RL-Gen and CN-Gen. We report in Table 6 the average quadratic Cohen's Kappa calculated over the three possible combinations of two judges (Toledo et al., 2019).

CN-Gen obtained the best scores during the automatic evaluation, while RL-Gen got the best scores from the human evaluation. The CN-Gen conversations are longer than the RL-Gen conversation by 1.3 turns on average (see Table 2), and hence would contain more word overlap with the ground truth. This results in better automatic evaluation scores for the CN-Gen, while the humans prefer short targeted conversations generated by RL-Gen.

3.2 Evaluation on the summarization task

To further evaluate the quality of the generate conversations, we augmented a conversation summarization dataset with generated conversations and evaluated the summarization model. We followed the following process: (1) We randomly selected x% of the summaries of the dataset and trained our conversation generation models, (2) The trained models were applied on the other (y=100-x%) of the summaries and generated conversations, (3) Those generated conversations along with the original summaries were added to the data. Using this approach, we can add extra y% (summary, conversation) pairs to the training data, (4) The conversation summarization model (discussed in Section 2 under 'Summary Generator') was trained on the augmented data. We compare the performance of the conversation summarization model on the original dataset and with augmentation.

Automatic Evaluation: We compare the three conversation generation methods at different augmentation percentages, and the results are shown in Table 7. At all augmentation levels, the summarization models trained with augmented data outperform the summarization model trained on the original dataset (without augmentation). CN-Gen based augmentation produces the best accuracy compared to other two methods. One prevalent pattern is that, when augmentation data increases, the accuracy seems to increase up to a certain point and then starts to decrease. The best accuracies were found around 30% data augmentation. We believe that more augmentation leads performance to drop due to the following reason. For augmenting with more data, we are left with less data to train the model for conversation generation (for 10% augmentation, the conversation generation models are trained on 90% of the data, while for 50% augmentation, the models are trained only on 50% of the

Model	Info	Gram	Cohe
Ground-Truth	4.56	4.46	4.47
SL-Gen	2.22	2.85	2.37
RL-Gen	**3.20**	**3.50**	**3.14**
CN-Gen	3.10	3.43	3.09

Table 5: Human evaluation of generated conversations

Model	Info	Gram	Cohe
Ground-Truth	0.04	0.22	0.25
SL-Gen	0.35	0.26	0.42
RL-Gen	0.47	0.35	0.45
CN-Gen	0.60	0.40	0.60

Table 6: Average Cohen's Kappa for human evaluation of generated conversations

Method	Augmentation %	ROUGE_1	ROUGE_2	ROUGE_L
	0% (Original)	51.84	30.98	43.98
SL-Gen	10%	52.82	31.99	44.89
	20%	**52.90**	**32.01**	44.97
	30%	52.88	32.02	**45.01**
	40%	52.61	31.98	44.96
	50%	52.55	31.98	44.80
RL-Gen	10%	52.93	32.05	44.92
	20%	53.30	32.15	45.20
	30%	**53.81**	**32.21**	**45.77**
	40%	52.86	32.06	44.99
	50%	52.64	32.07	44.88
CN-Gen	10%	53.29	32.36	45.08
	20%	53.36	32.53	45.27
	30%	**54.02**	**33.28**	**46.06**
	40%	52.14	31.76	44.14
	50%	52.36	31.75	44.85

Table 7: ROUGE F-1 evaluation on Samsum test set.

data). Therefore as the augmentation increases, the quality of generated conversations go down. This leads to overall smaller gains in the summarization task with increased augmentation after some point. To neutralize the effect of increasing the data points during augmentation, we experimented with a baseline which over-samples the original training data at different percentages to obtain same number of training instances as the augmented datasets. While the ROUGE-2 obtained with the original training data is 30.98, oversampling at 10%, 20%, 30%, 40% and 50%, only changes the ROUGE-2 to 30.55, 30.38, 30.74, 30.99 and 30.27 respectively. Hence, this suggests that oversampling hardly changes ROUGE scores obtained by training with the original dataset, while the augmentation according to our algorithms show significantly improved scores (as shown in Table 7).

Human Evaluation: We recruited 3 NLP experts to evaluate 50 instances of summaries generated with data augmentation (RL-Gen, CN-Gen), and respective summaries generated without augmentation (No-Aug). Here we consider two aspects with respect to a ground-truth summary: Coherency (whether the summary is easy to read) and Focus (whether the summary represents the ground-truth summary). Following (Amplayo and Lapata, 2020) we use the Best-Worst Scaling method. The score of each system is computed as the percentage of times it was chosen as the Best system minus times it was chosen as Worst. On the Coherency question, RL-Gen, CN-Gen and No-Aug obtained scores of 12.6, 6.6 and -4.0 respectively. On the Focus question RL-Gen, CN-Gen, and No-Aug obtained scores of 14.6, 6.0 and -2.6 respectively. These results confirm that the use of augmentation improves the quality of the summaries.

4 Conclusion

We investigated how the PLMs can be utilized to generate entire conversations that are grounded on a summary. We propose three approaches for conversation generation: SL-Gen, RL-Gen and CN-Gen and conducted multiple automatic and human evaluations to assess the quality of the generated conversations. Both automatic and human evaluations show that when compared to the ground truth conversations, RL-Gen and CN-Gen obtain high scores, suggesting that the proposed models generate high quality conversations. When a conversation summarization dataset is augmented with the generated conversations, the performance of conversation summarization is improved (over to 7% improvement in ROUGE-2 F-1), which also suggests that the proposed methods generate high quality conversations.

5 Ethics

We have used the publicly available Samsum dataset (https://huggingface.co/datasets/samsum). For the human evaluation of both conversations and summaries, we recruited 3 NLP researchers, who have graduate degree in NLP and Machine Learning. The annotation task itself was executed on Appen.com platform. Before the official annotation, we sampled 10 tasks to get an estimate of the duration of the task, and to make sure the instructions are clear enough.

References

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2-how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4766–4777.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2018. Towards automated customer support. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 48–59. Springer.

Yiping Kang, Yunqi Zhang, Jonathan K Kummerfeld, Lingjia Tang, and Jason Mars. 2018. Data collection for dialogue system: A startup perspective. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 33–40.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

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.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2020. Simulated chats for task-oriented dialog: Learning to generate conversations from instructions. *arXiv preprint arXiv:2010.10216*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

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.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A critical role dungeons and

dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. 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 5629–5639.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Qingyang Wu, Lei Li, and Zhou Yu. 2020. Textgail: Generative adversarial imitation learning for text generation. *arXiv preprint arXiv:2004.13796*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1008–1025.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

A Model Training and Hyperparameter Details

A.1 Supervised Conversation Generation (SL-Conv-Gen)

We fine-tune a GPT-2 language model using the implementation available at HuggingFace (Wolf et al., 2019). The hyper-parameters used during training and inference are shown below. The model takes around 6 hours to train on 2 V100 GPUs (single machine).

```
model_name_or_path: gpt2
per_gpu_train_batch_size: 4
per_gpu_eval_batch_size: 4
gradient_accumulation_steps: 4
learning_rate: 6.25e-5
adam_epsilon: 1e-8
max_grad_norm: 1.0
num_train_epochs: 10
warmup_steps: 500
min_length: 20
max_length: 512
top_k: 0
top_p: 0.95
```

A.2 Summary Generator

We use DistilBART instance[1] fine-tuned on the extreme summarization (XSum) task, and we fine-tune this model further on the Samsum dataset. The model takes around 12 hours to train on 2 V100 GPUs (single machine).

The hyperparameters used for training the DistilBART model are as follows:

```
train_batch_size: 4
eval_batch_size: 4
num_train_epochs: 10
model_name_or_path: sshleifer/distilbart
-xsum-12-6
learning_rate: 3e-5
val_check_interval: 0.1
max_source_length: 512
max_target_length: 80
```

A.3 Reinforced Learning based conversation generation (RL-Conv-Gen)

To train the RL based conversation generation model, we adapted a publicly available Proximal Policy Optimization (PPO) implementation [2]. The model takes around 12 hours to train on 2 V100 GPUs (single machine). Following hyperparameters were used to train the model.

```
steps: 10000
batch_size: 16
forward_batch_size: 4
learning_rate: 1.41e-5
init_kl_coef:0.2
target: 6
horizon:10000
gamma:1
lam:0.95
cliprange: 0.2
cliprange_value: 0.2
vf_coef: 0.1
```

[1] https://huggingface.co/sshleifer/distilbart-cnn-12-6

[2] https://github.com/lvwerra/trl

B Sample summaries with corresponding ground-truth

Figure 3 shows some samples of dialogs with their corresponding summaries - ground-truth and automatic generated ones.

Summary: Person0 closed some deals today. Person1 didn't manage to do it.

Ground Truth Dialog:	**SL-Gen:**	**RL-Gen:**	**CN-Gen:**
<Person0> It's been a long day. <Person1> Same here, Person0, good or bad? <Person0> Very busy. I closed some deals but I had a lot of stress also. <Person1> Lucky you. I couldn't close any deal. It annoys me. Some customers are negotiating for days, weeks. <Person0> Maybe they're just fishing. Want to compare offers. <Person1> That's what they do.	<Person0> Hi, are you still alive? <Person1> Hi, yes, still alive. <Person0> I closed some deals today. <Person1> Yeah, that's awesome. <Person0> Thanks.	<Person0> Hey, how are you? <Person1> Good, busy as usual. <Person0> I closed some deals today. <Person1> Too late, didn't manage to do it. <Person0> Sorry.	<Person0> Did you see this? ¡file_other¿ <Person1> Cool. What's up? <Person0> I closed some deals today and I feel good about it. <Person1> Great. I didn't manage to do it today. I'm sorry. I had no choice. <Person0> I'm sorry. <Person1> I didn't have the tools.

Summary: Person0 bought a table, six chairs, a vase and a pile of clothes and the second hand shop downtown. She paid 70 euros for everything.

Ground Truth Dialog:	**SL-Gen:**	**RL-Gen:**	**CN-Gen:**
<Person0> I went to the second hand shop downtown. <Person1> Cool Person0 What did you get? <Person0> Lots of stuff. A table, six chairs, a vase, a pile of clothes <Person1> That's really a lot of stuff Person0. Send us pictures <Person0> ¡photo_file¿ <Person1> Beautiful table.Is it wood? <Person0> Yes, oak. <Person1> Awesome! These are the clothes? I love the black dress. <Person0> It's absolutely beautiful. As soon as I saw it I knew it would be mine. I'm so happy with what I bought.	<Person0> I bought the second hand shop downtown. <Person1> So what happened? <Person0> I bought everything. <Person1> What do you mean? <Person0> I paid 70 euro for everything. <Person1> How much? <Person0> 6 chairs, six chairs, a vase and a pile of clothes. <Person1> Sounds like a bargain. <Person0> ou should be in business.	<Person0> Do you remember the second hand in downtown? <Person1> Yes, I do. What's inside? <Person0> A table, six chairs, a vase and a pile of clothes. <Person1> Is it expensive? <Person0> 70 euros.	<Person0> Look what I've bought! A table, six chairs, a vase. I bought a table, six chairs, a vase and a pile of clothes. <Person1> How much did you pay for everything? <Person0> I paid 70 euro for everything. And I thought it would cost me a fortune! I don't even know how much. <Person1> I am glad. <Person0> Me too. 70 euro for everything, including a vase!!!

Summary: Person1 is not at home. Person0 wants Person1 to keep her pasta in the microwave.

Ground Truth Dialog:	**SL-Gen:**	**RL-Gen:**	**CN-Gen:**
<Person0> U both at home? <Person1> No. I've just got ur msg. Why did u ask about it? <Person0> No reason. Keep my pasta in the microwave. <Person1> I haven't cooked anything.	<Person0> How are you doing? <Person1> I'm not at home <Person0> Why? <Person1> I wanted to keep some pasta in the microwave. <Person0> No need to. <Person1> In the microwave? <Person0> Yes.	<Person0> Are you at home? <Person1> No, I'm still at work. <Person0> Good, I should be home in about 15 minutes. <Person1> Can you put my pasta in the microwave? <Person0> Sure, I'll check it. <Person1> Thanks!	<Person0> Are you at home? <Person1> No, I'm busy at work. What do you want? <Person0> I want to put some pasta in microwave. <Person1> OK. I'll try.

Summary: Person0 needs Person1's help as he cannot get the application running.

Ground Truth Dialog:	**SL-Gen:**	**RL-Gen:**	**CN-Gen:**
<Person0> Person1, I need your help? <Person1> What's up, mate? <Person0> I can't get the application running. <Person1> Have you switched the computer on? <Person0> Very funny!!!!!!!!!!!!! <Person1> OK. Sorry. I can see it's serious. <Person0> Yeah, man. It is f**cking serious. <Person1> I'll be with you right now. <Person0> Thanks.	<Person0> Hey! I have a problem with the application. <Person1> Hello, Person0 <Person0> Could you please help me to get the application running? I cannot. <Person1> Why? <Person0> I have a problem with my laptop. The application is running on the USB port. I cannot access the computer. The computer is running. If I try to access the computer, the application will start. I hope that you can help me. <Person1> Okay. <Person0> Thank you, Person1.	<Person0> Person1, you there? <Person1> Yeah, what's up? <Person0> Do you need my help? <Person1> What's wrong? <Person0> I can't get the application running. <Person1> Did you try charging it first? <Person0> Yes, I tried that, but it doesn't work. <Person1> Try refreshing the database. <Person0> Damn, thanks!	<Person0> I cannot get the application running. <Person1> What's up? <Person0> Can you get the application running? <Person1> You sure you can't? Try and click on it. <Person0> You're so funny, Person1. <Person1> OK. I'm coming. <Person0> Thanks.

Summary: Person0 and Person1 will meet the new person in an hour.

Ground Truth Dialog:	**SL-Gen:**	**RL-Gen:**	**CN-Gen:**
<Person0> Can we meet with the new person soon? <Person1> Sure. In an hour okay? <Person0> Perfect.	<Person0> Hi, babe. <Person1> Hi. <Person0> We're meeting in an hour? <Person1> I'd prefer to be there in an hour. <Person0> Great!	<Person0> Did you get that message? <Person1> No, why? <Person0> I might have lost it. <Person1> Oh no, it's my mistake. I'm sorry. <Person0> Ok, don't worry. We can meet in an hour? <Person1> Ok!	<Person0> Wanna meet the new person? <Person1> Sure, I'll be there in an hour. <Person0> Perfect!

Figure 3: Samples of dialogs with their corresponding summaries - ground-truth and automatic generated ones