AutoConv: Automatically Generating Information-seeking Conversations with Large Language Models

Siheng Li[1†*], **Cheng Yang**[1†], **Yichun Yin**[2], **Xinyu Zhu**[1], **Zesen Cheng**[3]
Lifeng Shang[2], **Xin Jiang**[2], **Qun Liu**[2], **Yujiu Yang**[1‡]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Huawei Noah's Ark Lab, [3]Peking University
{lisiheng21, yangc21}@mails.tsinghua.edu.cn
{yinyichun, shang.lifeng, jiang.xin, qun.liu}@huawei.com
yang.yujiu@sz.tsinghua.edu.cn

Abstract

Information-seeking conversation, which aims to help users gather information through conversation, has achieved great progress in recent years. However, the research is still stymied by the scarcity of training data. To alleviate this problem, we propose AutoConv for synthetic conversation generation, which takes advantage of the few-shot learning ability and generation capacity of large language models (LLM). Specifically, we formulate the conversation generation problem as a language modeling task, then finetune an LLM with a few human conversations to capture the characteristics of the information-seeking process and use it for generating synthetic conversations with high quality. Experimental results on two frequently-used datasets verify that AutoConv has substantial improvements over strong baselines and alleviates the dependence on human annotation. In addition, we also provide several analysis studies to promote future research.

1 Introduction

In information-seeking conversations, users repeatedly ask questions based on their interests, and the dialogue system provides answers to fulfill their information needs (Stede and Schlangen, 2004; Choi et al., 2018; Reddy et al., 2019). This scenario is important for addressing real-world open-ended questions, which requires discussions to explore in depth (Dai et al., 2022), e.g., *How to learn more efficiently*? Though great progress has been achieved in recent years, most existing researches depend on abundant human annotation, which can be highly costly and limited in knowledge coverage.

A promising way to alleviate this problem is data augmentation (Chen et al., 2021). Traditional methods, including token-level manipulation (Kobayashi, 2018; Wei and Zou, 2019) and sentence-level paraphrasing (Sennrich et al., 2016), improve the linguistic diversity of training data. However, they cannot create conversations grounded on new documents, which are indispensable for dealing with out-of-domain scenarios. Another line of research focuses on simulation-based methods (Wu et al., 2021; Kim et al., 2022). Specifically, they can iteratively generate conversations grounded on new documents based on a span extractor and an utterance generator. Nevertheless, both the training of the extractor and the generator still require abundant human dialogues. Besides the above ways, Dai et al. (2022) propose Dialog Inpainting, which creates information-seeking dialogues by inserting utterances between neighboring sentences in documents. One potential risk is the gap between the structure of documents and that of conversations. Documents are tighter, while real-world conversations are more open-ended.

Method	DG	Data Needs
EDA (Wei and Zou, 2019)	✗	-
Back-Translation (Sennrich et al., 2016)	✗	-
SeemSeek (Kim et al., 2022)	✔	Large
Dialog Inpainting (Dai et al., 2022)	✔	Large
AutoConv (Ours)	✔	**Few**

Table 1: The differences between AutoConv and others. DG represents whether the augmentation is document grounded, and Data Needs denotes the scale of human conversations used for augmentation.

To alleviate the above issues, we propose a simple yet effective method **AutoConv** for **Auto**matically generating information-seeking **Conv**ersations, which takes advantage of the few-shot learning ability and generation capacity of large language models (LLM) (Brown et al., 2020). Specifically, we formulate conversation generation as a language modeling task and utilize an LLM for generating synthetic conversations grounded on external documents. Surprisingly, finetuning with a few human dialogues can help LLM capture the characteristics of the information-seeking process

* This work is done when Siheng Li is an intern at Huawei Noah's Ark Lab.
† Equal contribution.
‡ Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1751–1762

Figure 1: The generation process of AutoConv. We use nucleus sampling for generating user questions and greedy search for generating system answers.

(e.g., grounding, question answering) and generate high-quality synthetic conversations. Then, we can train a small task model with these dialogues. The differences between AutoConv and others are shown in Table 1.

We conduct comprehensive experiments on two frequently-used datasets QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019) in the low-resource setting, where only dozens of human dialogues are available. The results show that AutoConv has substantial improvements over several strong baselines. When scaling up the synthetic dialogues, AutoConv has the improvement of up to 5.06 F1 gain compared with directly finetuning, and thus largely reduces the labor force for annotation. In addition, we find that the small task model trained with synthetic dialogues can even surpass finetuned LLM with only 1.7% parameters. Moreover, we also investigate the impact of decoding strategy and scaling laws for AutoConv.

2 Method

2.1 Task Formulation

Our goal is automatically generating information-seeking conversations. Specifically, each conversation is grounded on a document $\boldsymbol{d}$ and consists of a series of user questions and system answers.

2.2 Conversation Generation

Training. We formulate conversation generation as a language modeling task and finetune[1] an LLM with a few human dialogues (e.g., 50 from QuAC (Choi et al., 2018)) to capture the characteristics of information-seeking conversations (e.g., grounding, question answering). The objective is the negative log-likelihood of each utterance:

$$\mathcal{L} = -\sum_{t=1}^{T}\sum_{l=1}^{L} \log P(u_l^t | u_{<l}^t, \boldsymbol{h}_{<t}, \boldsymbol{d}),$$

where $\boldsymbol{u}$ represents a user question or a system answer, $\boldsymbol{h}$ is the dialogue history, L and T are the number of tokens and turns respectively.

Generating. Based on the finetuned LLM, we can generate synthetic dialogues with unlabeled documents, as in Figure 1. In information-seeking scenarios, user questions are typically open-ended. Thus we choose nucleus sampling (Holtzman et al., 2020) for generating user questions, which has shown great performance in various open-ended generation tasks (Su et al., 2022). However, when applying a sampling decoding strategy for system answer generation, we find it results in the "hallucination" problem (Shuster et al., 2021), where the generation is plausible but factually incorrect based on the document. To this end, we utilize greedy search for answer generation. Neural language models often generate the same sentences repetitively (Xu et al., 2022). To alleviate this problem, we first compute the diversity score of each synthetic dialogue as in Su et al. (2022), which considers the repetition at different n-gram levels. Then, we filter out dialogues based on this score.

After that, a two-stage training strategy is adopted (Xie et al., 2020b) for training a small task model. Specifically, we first pre-train it on the synthetic dialogues, then finetune it on the human dialogues used for finetuning the LLM. More training details are given in Appendix B.

3 Experiments

We conduct experiments on QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019), more details about them are shown in Appendix A.

3.1 Implementation

We focus on the low-resource setting, where human dialogues are scarce. To simulate this setting, we randomly sample a few human dialogues from the training set of QuAC or CoQA, and use them for finetuning the LLM. We use OPT-13B (Zhang et al., 2022) as the LLM and UnifiedQA-V2-base (222M) (Khashabi et al., 2022) as the small task model. All data augmentation methods use the same training strategy and small task model. More implementation details are shown in Appendix B.

3.2 Comparison with Baselines

We compare AutoConv with a series of baselines, and the details of them are given in Appendix C. As

[1] In our preliminary experiments, we try to prompt LLM without training. However, we find that the performance is poor and LLM fails to generate conversations grounded on the documents, similar to the observation in Zheng et al. (2022).

Method	QuAC		CoQA	
	F1	EM	F1	EM
Prompting				
GPT-3 Zero-shot (Brown et al., 2020)	41.5	-	81.5	-
GPT-3 Few-shot (Brown et al., 2020)	44.3	-	85.0	-
Data Augmentation (50 Human Dialogues)				
Finetuning	46.57±1.29	30.68±1.25	70.41±0.46	60.43±0.56
Back-Translation (Sennrich et al., 2016)	47.92±0.49	28.26±1.39	67.59±2.73	56.34±3.41
EDA (Wei and Zou, 2019)	46.04±1.28	28.88±2.20	58.89±2.08	47.64±2.14
Utterance Manipulation (Chen and Yang, 2021)	48.83±0.63	33.91±0.73	68.69±0.85	58.30±1.21
Dialog Inpainting (Dai et al., 2022)	48.33±1.24	32.23±1.55	70.25±0.93	59.83±0.98
AutoConv	**50.48**±0.94	**34.12**±0.93	**73.87**±0.85	**63.78**±1.01
Human Annotation	53.24±0.28	36.85±0.35	76.02±0.71	65.92±1.01
Data Augmentation (100 Human Dialogues)				
Finetuning	48.98±1.16	31.98±1.09	72.78±0.69	62.41±0.85
Back-Translation (Sennrich et al., 2016)	48.41±0.96	28.10±2.51	69.18±2.82	57.72±3.28
EDA (Wei and Zou, 2019)	46.86±0.61	29.14±1.71	60.61±4.23	49.24±4.74
Utterance Manipulation (Chen and Yang, 2021)	49.07±1.06	31.77±1.86	69.23±0.21	59.15±0.74
Dialog Inpainting (Dai et al., 2022)	49.48±0.34	33.29±0.98	72.15±0.74	61.80±0.99
AutoConv	**51.21**±1.02	**34.65**±1.00	**74.84**±0.24	**64.36**±0.46
Human Annotation	54.22±0.90	37.42±2.06	76.35±0.51	65.71±0.55

Table 2: Comparison with baselines. All experiments are performed 4 runs with different random seeds. Finetuning means directly training with only human dialogues. All data augmentation methods use the same human dialogues and the same number of synthetic dialogues for the sake of fairness (5 times the number of human dialogues). Human annotation represents replacing the synthetic dialogues with the same number of human dialogues.

shown in Table 2, AutoConv achieves better performance than GPT-3 prompting on QuAC with only 0.13% parameters and 50 human dialogues, but is less competitive on CoQA. We conjecture the reason stems from the intrinsic difference between the two datasets. CoQA contains more factoid questions, and the answers are named entities or short noun phrases like those in SQuAD (Rajpurkar et al., 2016). By training on large-scale text corpus from a web forum, GPT-3 might implicitly learn the format and structure of question answering (Sanh et al., 2022), and thus gets excellent performance on CoQA. On the other side, QuAC has more open-ended and exploratory questions as in natural conversations, and 86% questions are contextual (Choi et al., 2018). Therefore, it brings more difficulties for GPT-3 inference with few demonstrations, while our method learns better from both human dialogues and synthetic dialogues.

Compared with data augmentation methods, AutoConv achieves the best performance on both datasets and mitigates the gap between synthetic dialogues and human upper bounds. We find that the token-level augmentation method EDA and the sentence-level augmentation method Back-Translation even hurt the performance, which is similar to the observation in Chen et al. (2021). One possible reason is that they bring too much noise. Dialog Inpainting (Dai et al., 2022) gets ordinary performance, and the reason possibly derives from the gap between the structure of natural conversations and that of the documents used for constructing synthetic dialogues.

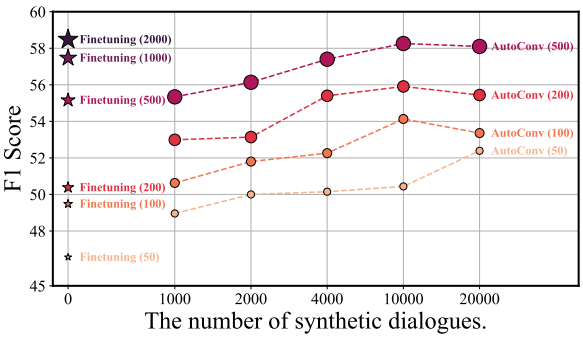

Figure 2: The results of scaling up human dialogues and synthetic dialogues on QuAC. The number in the parentheses represents the number of human dialogues.

3.3 Scaling up Human Dialogues and Synthetic Dialogues

In this part, we further analyze the performance of AutoConv when scaling up the human dialogues and synthetic dialogues. As shown in Figure 2, the

Model	#Params	#FLOPs	F1 (50)	F1 (200)
Finetuning (LLM)	12.9B	7049.3B	**53.53**	54.85
Finetuning (STM)	**222M**	**60.2B**	47.97	50.38
AutoConv (STM)	**222M**	**60.2B**	52.40	**55.44**

Table 3: Comparison results on QuAC. Finetuning means training with only human dialogues. AutoConv uses the same human dialogues and 20K synthetic dialogues. LLM is large language model and STM is small task model. The number in the parentheses represents the number of human dialogues.

performance boosts when more human dialogues or synthetic dialogues are used. With 50 human dialogues, AutoConv outperforms the results of finetuning with 200 human dialogues. With 500 human dialogues, AutoConv gets competitive performance compared with finetuning with 2000 human dialogues. These results verify the high quality of synthetic dialogues, and our AutoConv can largely alleviate the labor force for annotation.

3.4 Comparison with Finetuned Large Language Model

AutoConv is a kind of symbolic knowledge distillation (West et al., 2022), where the finetuned large language model (LLM) transfers its knowledge to the small task model (STM) by generating synthetic dialogues for the training of STM. Here, we further investigate the effectiveness of AutoConv from the aspect of knowledge distillation. As shown in Table 3, finetuned LLM has substantial improvements over finetuned STM. However, it brings large memory and computation cost. On the other side, our AutoConv not only keeps the efficiency of STM, but also boosts the performance. Surprisingly, AutoConv even outperforms its teacher model in the 200 human dialogues setting. Similar observations are found in West et al. (2022); Ye et al. (2022), while they focus on different tasks. We leave the analysis of this novel observation for future work.

3.5 Impact of Decoding Strategy

During our preliminary experiments, we find that the decoding strategy is important for system answer generation. More precisely, we evaluate the answer generation performance of LLM with different decoding strategies on QuAC, and the results are shown in Table 4. Though nucleus sampling (Holtzman et al., 2020) has shown great performance in various generation tasks (Su et al., 2022), it performs less competitively than maximization-based decoding strategies for answer generation. Compared with beam search, greedy search shows competitive performance and is more efficient. Thus we use greedy search by default in this paper.

Decoding Strategy	F1	Exact Match
Nucleus Sampling ($p = 0.8$)	50.77	32.63
Nucleus Sampling ($p = 0.9$)	49.88	31.57
Greedy Search	53.53	36.38
Beam Search ($b = 4$)	**54.43**	38.64
Beam Search ($b = 8$)	**54.43**	**38.70**

Table 4: The results of LLM with different decoding strategies for answer generation on QuAC, 50 human dialogues are used for finetuning the LLM.

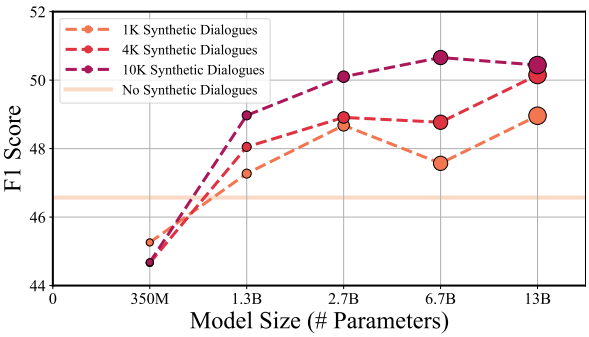

Figure 3: The results of AutoConv with different LLM on QuAC. We use different scale of OPT (Zhang et al., 2022) as the LLM. All models are trained with 50 human dialogues for fairness and synthetic dialogues are generated with the corresponding LLM.

3.6 Scaling Laws

We further analyze how the benefit of AutoConv is affected by the scale of LLM. As shown in Figure 3, the performance gets better with a larger model across a various number of synthetic dialogues. In addition, when the LM is small (350M) and with limited generation ability, the synthetic dialogues can even hurt the performance when the available human dialogues are scarce. Due to the limitation of computational resources, we limit our investigation to 13B parameters and leave larger models for future work.

3.7 Case Study

In Table 5, we present an example of our synthetic conversation for the case study. The original document describes the singer Ciara's second studio album and her acting debut. The conversation consists of seven user questions and seven system answers, covering the title and sales of the album, the duration of the tour, etc. As we can see from this

Title
2006-2007: Ciara: The Evolution and acting debut
Document
On December 5, 2006, Ciara released her second studio album, Ciara: The Evolution. According to the singer, the title of the album is \"about so much more than just my personal growth - it's about the evolution of music, the evolution of dance, the evolution of fashion \" The source of the albumś creativity such as the sound and edge comes from Ciara in general. Ciara: The Evolution became Ciaraś first and only number one album on the U.S. Billboard 200, and her second number one on the Top R&B/Hip-Hop Albums charts with sales of 338,000 in the first week . . .
Dialogue
Usr: What was the evolution?
Sys: On December 5, 2006, Ciara released her second studio album, Ciara: The Evolution.
Usr: What was the name of the album?
Sys: The title of the album is \"about so much more than just my personal growth - it's about the evolution of music, the evolution of dance, the evolution of fashion.\"
Usr: Did the album do well?
Sys: The album became Ciara's first number one album on the U.S. Billboard 200 and second on the Top R&B/Hip-Hop Albums charts with sales of more than 338,000.
Usr: What awards did she win if any?
Sys: The album went on to be certified platinum by the RIAA in the United States, within only five weeks of its release, and has sold 1.3 million copies according to Nielsen SoundScan.
Usr: What else did she accomplish during that time?
Sys: In support of the album, Ciara went on her first headlining tour in October 2006.
Usr: How long was the tour?
Sys: The tour went to seventeen different clubs in cities throughout the United States.
Usr: Did she have success with any songs?
Sys: The lead single, \"Get Up\", which features Chamillionaire, reached number seven in the United States and gained a platinum accreditation.

Table 5: An example of the synthetic conversation generated by AutoConv, the LLM is finetuned with 50 human dialogues from QuAC (Choi et al., 2018).

example, the user questions are diverse (e.g. what, how, did, etc.) and the conversation is informative and conversational. For example, when the system mentions "tour" (the fifth system utterance), the user follows by asking "How long was the tour?".

3.8 Error Analysis

To further analyze the limitation of our method, we conduct an error analysis by manually investigating 50 synthetic conversations generated by AutoConv, which is finetuned with 50 human conversations from QuAC (Choi et al., 2018). Particularly, we find that only 5% generated questions are not suitable (e.g., misspelled names). The reason stems from the open-ended characteristic of natural conversation that many kinds of user questions are possible under the same context. However, nearly 40% of system answers are not perfect, and we summarize the wrong answers into four major classes: **(1) Irrelevant**: 75% of them are totally irrelevant to user questions. **(2) Related but not Accurate**: 14% of them contain related knowledge from the grounded documents, but the answers are not accurate. Take an example in Table 5, the second user question asks for the name of the album, which is *Ciara: The Evolution* according to the document. While the LLM generates the interpretation of the album name by mistake. **(3) Missing**: 4% of them belong to the missing error that the system answers are "No Answer", while the questions actually can be answered based on the documents. **(4) Hallucination**: 3% of them mention hallucination knowledge, which cannot be found in the documents. In addition, we also notice that AutoConv is more likely to generate wrong answers when grounding on longer and more complex documents.

4 Conclusion

In this paper, we propose a simple yet effective method, AutoConv, which formulates the conversation generation problem as a language modeling task. Then, based on a large language model and a few human dialogues, AutoConv can generate synthetic dialogues with high quality. Experimental results on both QuAC and CoQA verify the effectiveness of AutoConv, which alleviates the human efforts for annotation largely. Furthermore, we also provide case study and error analysis to prompt future research.

Limitations

In this paper, we propose a method named AutoConv, which means automatically generating information-seeking conversations with large language models (LLM). Though it has achieved great performance on both QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019), there are still some limitations that should be noticed.

Limitation of LLM. In our experiments, we use OPT-13B (Zhang et al., 2022) as the LLM for generating synthetic conversations due to the limited computational resources. Larger models should be considered to further understand the potential ability of AutoConv, e.g., GPT-3 (Brown et al., 2020), OPT-175B (Zhang et al., 2022), BLOOM-176B (Scao et al., 2022), and GLM-130B (Zeng et al., 2022) etc.

Limitation of Implementation. As mentioned in Section 2.2 and Appendix B, our method needs to finetune LLM and generate massive synthetic conversations based on the finetuned LLM, which has a high cost for implementation.

Limitation of Synthetic Dialogues. As shown in Table 2 and Section 3.8, there is still a gap between our synthetic dialogues and human dialogues. It is important to improve the quality of synthetic dialogues so that we can further alleviate the dependence on human annotation.

Ethics Statement

AutoConv is based on large language models (LLM), while LLM has some potential risks, e.g., social bias (Liang et al., 2021), offensive content (Ganguli et al., 2022) etc. Fortunately, we finetune the LLM to capture the characteristics of the information-seeking process, and the generated conversations are mostly grounded on the provided documents (take an example in Table 5). Therefore, our method alleviates the potential risks of directly using LLM. According to our manual check in error analysis (Section 3.8), we do not find any harmful content in the synthetic conversations. In addition, we also encourage considering more safety methods (Xu et al., 2020; Sun et al., 2022) to guarantee the quality of synthetic conversations.

Acknowledgements

This work was partly supported by the National Key Research and Development Program of China (No. 2020YFB1708200) , the "Graph Neural Network Project" of Ping An Technology (Shenzhen) Co., Ltd. and AMiner.Shenzhen SciBrain fund.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in NLP. *CoRR*, abs/2106.07499.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6605–6616. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson,

Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *CoRR*, abs/2209.07858.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *CoRR*, abs/2202.12359.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1896–1907. Association for Computational Linguistics.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Towards more realistic generation of information-seeking conversations. *CoRR*, abs/2205.12609.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 452–457. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings*

of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pages 3784–3803. Association for Computational Linguistics.

Manfred Stede and David Schlangen. 2004. Information-seeking chat: Dialogues driven by topic-structure. In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*. Citeseer.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *CoRR*, abs/2202.06417.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3906–3923. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Qingyang Wu, Song Feng, Derek Chen, Sachindra Joshi, Luis A. Lastras, and Zhou Yu. 2021. DG2: data augmentation through document grounded dialogue generation. *CoRR*, abs/2112.08342.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020a. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020b. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. Computer Vision Foundation / IEEE.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *CoRR*, abs/2206.02369.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *CoRR*, abs/2010.07079.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *CoRR*, abs/2202.07922.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. GLM-130B: an open bilingual pre-trained model. *CoRR*, abs/2210.02414.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *CoRR*, abs/2202.13047.

A Datasets

QuAC. QuAC (Choi et al., 2018) is a leading conversational question answering dataset, consists of 14K information-seeking dialogues. Different from the factoid questions in most existing QA datasets, the questions in QuAC are more open-ended and exploratory. In addition, 86% of questions are contextual, and the model needs to understand the dialogue context to resolve coreference. As the test set is only available in the QuAC challenge[2], we evaluate the performance on the development set.

CoQA. CoQA (Reddy et al., 2019) consists of 127K conversational QA pairs across seven domains. Different from QuAC, CoQA focus more on factoid questions, and the answers are mostly named entities or short phrases as in SQuAD (Rajpurkar et al., 2016). The test set of CoQA is only available in the CoQA challenge[3], therefore we evaluate the performance on the development set.

B Implementation Details

General Setting. All experiments are based on Transformers[4] (Wolf et al., 2020), DeepSpeed[5] (Rasley et al., 2020) and Pytorch Lightning[6]. We use UnifiedQA-V2-base[7] (Khashabi et al., 2020, 2022) as the small task model, which is based on T5 architecture with 222M parameters and pre-trained on many QA tasks (the tasks in our experiments are not included in). The training of the small task model follows the original paper (Khashabi et al., 2020) in a Text-to-Text framework (Raffel et al., 2020). The input is `Dialogue History \n Document` and the output is `System Answer`.

For the training hyperparameters, we set the learning rate as $3e-4$, batch size as 32, and use Adam optimizer (Kingma and Ba, 2015) with warmup learning rate schedule, the warmup ratio is 0.1. When comparing with baseline methods as in Section 3.2, all methods use the same small task model, the same two-stage training strategy (Xie et al., 2020b; Chen and Yang, 2021), the same human dialogues and the same number of synthetic dialogues for fairness (5 times the number of human dialogues). For the 50 human dialogues setting, we train each model for 1K gradient steps in the pre-training stage and 200 gradient steps in the fintuning stage. For the 100 human dialogues setting, the steps are 2K and 400 respectively. When scaling up the number of synthetic dialogues as in Section 3.3 and Section 3.6, the numbers of pre-training steps scale up, which are 2K, 4K, 8K, 20K and 40K for 1K, 2K, 4K, 10K and 20K synthetic dialogues respectively, and the finetuning steps are 200, 400, 800 and 2K for 50, 100, 200 and 500 human dialogues respectively. For all experiments, we randomly sample 20% dialogues as the validation set, and others as the training set. The model is validated every epoch, and we choose the checkpoint with the best F1 score on the validation set for evaluation.

Ours. We use OPT-13B[8] (Zhang et al., 2022) as the LLM for generating synthetic dialogues, which is a decoder-only pre-trained language model with 13B parameters. The learning rate and batch size are set as 1e-5 and 32. Adam optimizer (Kingma and Ba, 2015) with warmup learning rate schedule is utilized for optimization and the warmup ratio is 0.1. The max training steps of LLM are 200, 400, 800 and 2K for 50, 100, 200 and 500 human dialogues respectively. According to the performance of AutoConv on the validation set of human dialogues, we find that training LLM for 4 epochs is the most suitable. We randomly sample 5K documents from the training sets of QuAC and CoQA, and generate 8 synthetic dialogues for each document. The number of turn is set as 14 for QuAC and 30 for CoQA. Then, we filter a quarter of the synthetic dialogues based on the diversity score of each dialogue as in Su et al. (2022), which takes into account the repetition at different n-gram levels. It takes around 5 hours for training LLM and 18 hours for generating synthetic dialogues with 8 Tesla V100 32GB GPUs.

Evaluation. To evaluate the quality of synthetic conversations, we evaluate the conversational question answering performance of the small task model, which is trained on both synthetic conversations and a few human conversations. The metrics are Exact Match and word-level F1 as in Choi et al. (2018).

[2] https://quac.ai/

[3] https://stanfordnlp.github.io/coqa/

[4] https://huggingface.co/docs/transformers/index

[5] https://github.com/microsoft/DeepSpeed

[6] https://github.com/Lightning-AI/lightning

[7] https://huggingface.co/allenai/unifiedqa-v2-t5-base-1363200

[8] https://huggingface.co/facebook/opt-13b

C Baselines

Prompting. Prompting is a promising method for many NLP tasks. It aims to elicit the ability of large language models learned from pre-training with text demonstrations (e.g., task instruction and few-shot examples etc). In Table 2, we report the results from Brown et al. (2020).

Finetuning. Train the small task model with only human annotations.

EDA. Easy Data Augmentation (EDA) is a simple but effective method for text classification (Wei and Zou, 2019). Given an input text, including both the knowledge paragraph and dialogue history in our experiments, four operations are applied to create new examples, including synonym replacement, random insertion, random swap and random deletion. We use their open source code[9] for implementation.

Back-Translation. Back-Translation is one of the most popular augmentation method for NLP tasks (Sennrich et al., 2016; Xie et al., 2020a). Specifically, we first translate the input text to a target language, then translate it back to the source language, thus we can get a paraphrased example. To get various augmentations for each sample, we use five target languages, including Chinese, French, German, Arabic, and Korean. Huawei Translate[10] is used for the translation process.

Utterance Manipulation. Chen and Yang (2021) propose utterance-level manipulation to perturb the discourse relations in the conversation. Two simple operations are used: (1) random swapping, which randomly swaps two utterances to mess up the logic chain of the conversation, and (2) random deletion, which means randomly deleting an utterance to improve the discourse diversity. We randomly select one operation for each augmentation.

Dialog Inpainting. The state-of-the-art data augmentation method for conversational question answering. Given a document, they iteratively insert generated utterances between the consecutive sentences in the document, then the utterances and sentences can form an informative conversation (Dai et al., 2022). We randomly sample generated dialogues from their open source data[11].

[9] https://github.com/jasonwei20/eda_nlp

[10] https://www.huaweicloud.com/product/nlpmt.html

[11] https://github.com/google-research/dialog-inpainting

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction

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

B ☑ Did you use or create scientific artifacts?

Appendix A and Appendix B

☑ B1. Did you cite the creators of artifacts you used?
Appendix A

☐ 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)?
Ethics Statement and Appendix A

☐ 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.?
Appendix A

☑ 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.
Appendix A and Appendix B

C ☑ Did you run computational experiments?

Section 3

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

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?
Appendix B

☑ 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 3

☑ 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.)?
Appendix B

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

Left blank.

☐ 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.?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ 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.