AUGESC: Large-scale Data Augmentation for Emotional Support Conversation with Pre-trained Language Models

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Minlie Huang
The CoAI Group, DCST, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Abstract

Crowd-sourcing is commonly adopted for dialog data collection. However, it is highly costly and time-consuming, and the collected data is limited in scale and topic coverage. In this paper, aiming to generate emotional support conversations, we propose exploiting large-scale pre-trained language models for data augmentation, and provide key findings in our pilot exploration. Our adopted approach leverages the 6B-parameter GPT-J model and utilizes publicly available dialog posts to trigger conversations on various topics. Then we construct AUGESC[1], a machine-augmented dataset for emotional support conversation. It is two orders of magnitude larger than the original ESConv dataset in scale, covers more diverse topics, and is shown to be of high quality by human evaluation. Lastly, we demonstrate with interactive evaluation that AUGESC can further enhance dialog models tuned on ESConv to handle various conversation topics and to provide significantly more effective emotional support.

1 Introduction

In recent years, data-driven approaches have taken the power of dialog models to unprecedented heights (Zhang et al., 2020; Bao et al., 2021; Roller et al., 2021), which are able to produce engaging, informative, and human-like responses. To endow dialog models with more sophisticated capabilities, such as emotional support (Liu et al., 2021), researchers usually resort to crowd-sourcing to collect artificial dialog data, because the corpora required for many tasks do not naturally exist.

However, crowd-sourcing has several drawbacks. First, it is highly costly and time-consuming, especially when essential worker training and data qualification screening are taken into consideration (Liu et al., 2021). Second, due to budget constraints, the collected conversations are limited in scale and topic coverage. As a result, dialog models trained on such corpora may have trouble in dealing with out-of-domain topics (i.e., topics not appearing in training data), as shown in Figure 1.

[1]Data will be made accessible after paper publication.

User: I feel so frustrated. _My coworker is allowed to work remotely, but I am not..._
System: I can understand that, it is hard to ***be around people you don't know***.
User: I don't understand _why my boss treats us differently_. I also want to work remotely
System: ***When did you start working? Do you have a family?***
User: I joined the company three years ago and I'm married with two children. _So I want to spend more time with them by working remotely_
System: Ah, I see. You are wanting to be able to spend time with your family and ***not just be around your co-workers***. I can definitely understand that.

Figure 1: A pilot interaction example with Blender fine-tuned on ESConv (Liu et al., 2021). ***Blue texts*** reveal that Blender does not understand the user's situation and causes of frustration, despite that the user provides clues continuously (_underlined texts_).

In this paper, we target the task of Emotional Support Conversation (ESC) (Liu et al., 2021), whose original corpus ESConv has only 1.3K dialog sessions and covers very limited topics. To address the above challenges, we propose exploiting large-scale pre-trained language models for automatic dialog data augmentation. Our contributions fall into the following three aspects.

First, we summarize several key findings in our pilot exploration of methodology (§3). Inspired by these findings, our adopted approach (§4) leverages the currently largest open-sourced GPT model, GPT-J, and utilizes publicly available dialog posts to trigger conversations on various topics.

Second, we construct AUGESC, a large-scale machine-augmented dataset for ESC (§5). It is two orders of magnitude larger than the original ESConv dataset in scale and has wider and more diverse topic coverage. We provide detailed data statistics and analysis, and show the high quality of AUGESC through human evaluation (§6).

Third, interactive evaluation (§7) demonstrates that AugESC can further enhance dialog models tuned on ESConv to handle various conversation topics and to provide significantly more effective emotional support.

2 Related Work

Pre-trained Models The emergence of pre-trained models (PMs) has greatly promoted the progress of NLP technology in the past few years (Devlin et al., 2019; Radford et al., 2019). A PM is commonly trained on massive textual data with unsupervised learning objectives to capture general language features. For instance, DialoGPT (Zhang et al., 2020) and PLATO-2 (Bao et al., 2021) acquire the ability of open-domain conversation through pre-training on huge amounts of Reddit data. Blender (Roller et al., 2021) further establishes more attractive traits, such as being knowledgeable and empathetic, after being fine-tuned on more focused downstream tasks (Zhang et al., 2018; Dinan et al., 2018; Rashkin et al., 2019).

Data Augmentation with PMs PMs also bring opportunities for automatic data augmentation of various NLP tasks. Schick and Schütze (2021) combine the GPT-2 model (Radford et al., 2019) and task instructions to automatically generate a *textual similarity* corpus. Wang et al. (2021) utilize the powerful gigantic GPT-3 model (Brown et al., 2020) for data augmentation of *text classification and language understanding* tasks, while West et al. (2021) leverage GPT-3 to acquire *commonsense knowledge graphs*. Our work is distinguished from the above ones in that we focus on *dialog* data augmentation, which is more complex and challenging due to the much longer data length and the higher requirements for data quality.

A more relevant work to ours is (Mohapatra et al., 2020). They use the GPT-2 model to simulate the interaction between crowd-sourcing workers by training a user bot and an agent bot. Our work is distinct from theirs in two aspects. (1) Instead of simulating real interaction with chats between two role-related bots, we treat dialog data augmentation as a language modeling task (Figure 2). (2) Beyond verifying the approach's effectiveness in training end-task models (§7), our work also provides valuable insights on the approach selection for dialog data augmentation (§3), and concentrates more on improving, analyzing and evaluating the quality and utility of the augmented data (§4, 5, 6).

Emotional Support Conversation (ESC) ESC is a novel dialog task proposed recently (Liu et al., 2021), in which the help-seeker (or the user) is in a negative emotional state and the supporter (or the system) has the goal of providing emotional support to reduce the user's emotional distress. The authors crowd-sourced a dataset (namely ESConv), which contains 1.3K dialog sessions and 13 seekers' problem types. However, the construction of ESConv includes elaborate worker training and quality control mechanisms, which greatly restricts the scale and topic coverage of ESConv. Our work aims at the ESC task due to its significant application value and the intractability of collecting high-quality dialog data. Note that our approach is also applicable to other dialog tasks, such as knowledge-grounded dialog (Dinan et al., 2018), which can be studied in future research.

3 Key Findings

In our pilot exploration, we conducted extensive experiments to investigate feasible solutions for dialog data augmentation[2]. Our exploration mainly aimed to answer the following questions: (1) Should we select conversational models (e.g., Blender and DialoGPT) or autoregressive language models (i.e., the GPT models, such as GPT-3) to simulate interlocutors' interaction or generate conversations through language modeling? (2) How to fine-tune or prompt models to produce complete and coherent conversations? (3) How to trigger informative and diverse conversations? Our key findings are summarized as follows.

1. Language models are superior to conversational models (§6.3). (1) Language models store much richer knowledge learned from massive training corpora, which facilitates better generalization to various conversation topics. (2) We observed that GPT-generated conversations have much better dialog coherence and consistency than those simulated by conversational models (e.g., Blender).

2. Language modeling is more suitable for open-ended dialog data augmentation than interactive simulation. Without external constraints (e.g., user goals and knowledge bases in task-oriented dialog, Mohapatra et al. 2020), we observed little difference between conversations generated by one GPT model and those simulated by two separate models. However, the former is

[2]Our work assumes that the original dialog corpus (e.g., ESConv) can be accessed to assist data augmentation.

The following is a conversation with an AI assistant. The assistant is helpful, empathetic, clever, and very friendly. It can use various support skills to provide emotional support to human.

Human: I moved into a new state recently, and there's a lot to do, but I don't have any friends in the new place I stay at.
AI: What's it like being away from family?
Human: Family is all I have here. They aren't exactly close, so I haven't gotten to see them in a while.
AI: That must be difficult. How do you feel about where you live?
Human: It's OK. I'm learning to like it a little bit. At least now I have someone who is usually around when I wake up.
AI: If only you were that lucky with people in general. People move for so many different reasons. I've found that often when I move, I just need to adjust my social circle a little, and I find that I then end up liking where I am.
Human: That's true. Maybe I should just find some people to hang out with.
AI: Well, of course! There are a LOT of online social groups. We can start with that.
Human: I think that's a good idea. And I already know one person at my current location.
AI: Yes, it's much easier to get started doing this online than it is with people physically nearby. I know that you were actually just thinking that. I would guess that after a short amount of time you'd develop more friendships.
Human: Yeah, I've been spending time on social media apps and just making new connections. I've found some really good friends through it.
AI: That's wonderful! I'd say, you're on your way!
Human: What should I do first?
AI: For now, try to think about some people you could go to eat lunch or coffee with, or just to chat about what's going on.
Human: Oh yeah, that's a great idea! Thank you!
AI: You are welcome! Of course, when you can afford it, do get together with others in real life! It'll be much more rewarding than anything you could do alone.
Human: Do you think that will help? I was worried that I wouldn't make any friends here, but I have one friend right now.
AI: It certainly will help immensely. Real life friends are the best of all!
Human: That's good to hear! I guess I'll keep trying. Thanks for the help!
AI: You are welcome! And again, I wish you the best of luck with your new situation!

Figure 2: Illustration of our approach and a cherry-picked example conversation from AUGESC. An autoregressive language model (our approach adopts GPT-J) is first given the template prompts (highlighted texts), which contain a task instruction (the first paragraph), a trigger query (the first dialog post starting with "Human: "), and the next "AI: ". The language model then generates the subsequent conversation (non-highlighted texts) until the `[EOS]` token is decoded. All line breaks in the conversation are represented by `\n` in the input text.

greatly superior to the latter in terms of computational efficiency.

3. Prompted GPT-3 is inferior to fine-tuned GPT models. (1) Prompted GPT-3 has poor controllability of generation and usually fails to produce valid dialog data (§4.3). (2) Due to the lack of fine-tuning, GPT-3 has difficulty in grasping the task specifications (e.g., the conversation scenario) and the characteristics of our required data (e.g., the conversation should be empathetic and supportive). As a result, its generated conversations show significantly lower quality than those generated by fine-tuned models (§6.3).

4. Few-shot fine-tuning leads to better generalization and higher diversity (§4.1). In the early exploration, we found that fine-tuning with a few samples and training steps can greatly maintain the intrinsic knowledge of language models. In contrast, increasing tuning samples or training steps would lead to poor generalization to out-of-domain topics and failure to trigger the originally intrinsic knowledge. This finding is interesting and important as it indicates that with the help of large-scale automatic data augmentation, training dialog models may only need a few manually curated dialog samples.

5. Informative queries (the first dialog posts) are necessary for triggering on-topic conversations (§4.2). Generic and uninformative queries tend to result in off-topic and shallow conversations.

4 Methodology

Our approach treats dialog data augmentation as a language modeling task. Figure 2 illustrates how our approach works. The rest of this section elaborates on the details of our approach.

4.1 Implementation Details

Backbone Model We use the currently largest open-sourced GPT model, GPT-J (Wang and Komatsuzaki, 2021), as our backbone model. GPT-J has 6B parameters in total, accepts the maximum input length of 2,048, and is pre-trained on the 800GB Pile corpus (Gao et al., 2020).

Template Prompts As shown in previous research (Zheng and Huang, 2021), template prompts facilitate the performance of zero- or few-shot generation of language models. Thus, we first prepend the model input with a task instruction, as shown in

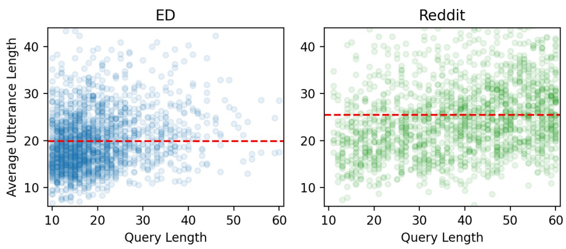

Figure 3: Relation between the query length and the average length of utterances in the triggered conversation. ***Red*** horizontal dashed lines denote the mean values of average utterance lengths. Statistics for each set are based on randomly sampled 2K GPT-J-generated conversations *before any filtering*. Queries from Reddit (***right***) are usually longer than those from ED (***left***), leading to longer utterances in triggered conversations.

Figure 2, which defines (1) a *conversation* scenario between a *human* and an *AI assistant*, and (2) that *the AI assistant can use various support skills to provide emotional support to human*. Then given the first dialog post as the trigger query, the model generates the subsequent conversation. Both GPT-J and GPT-3 use the same template prompt.

Note that unlike (Zheng and Huang, 2021), we do not insert descriptions of support strategies (defined in the ESC paper) before AI's utterances. This is because we found that the original strategy annotation is noisy, and inserting strategy descriptions also makes the generation less controllable.

Training We implemented GPT-J using the Transformers library (Wolf et al., 2020). We fine-tuned GPT-J for 1 epoch on 100 ESConv dialog sessions, which are randomly sampled and cover the 13 problem types evenly. The batch size was set to 2, and the language modeling loss was averaged over all the tokens in conversations (excluding the task instruction texts). We used the AdamW optimizer (Loshchilov and Hutter, 2018), the learning rate 6e-5, and the linear learning rate scheduler with warmup steps 5. We set the maximum input length to 1,500, and applied gradient checkpointing and model parallelism to reduce GPU memory occupation. Our fine-tuning of GPT-J requires at least 4 Tesla V100 32GB GPUs.

Inference The maximum generation length was also set to 1,500. We adopted Top-p sampling (Holtzman et al., 2019) with p set to 0.9 and temperature to 0.9. We set the repetition penalty factor to 1.05 to avoid generation of duplicate contents.

4.2 Trigger Queries

To trigger conversations on various topics, we collected informative dialog posts of two sources as queries: EmpatheticDialogues (ED) (Rashkin et al., 2019) and Reddit. ED is a crowd-sourced empathetic dialog dataset, where the dialog posts are assigned with the corresponding emotion labels. We found that most of the dialog posts in ED contain detailed descriptions about the emotional states, and we thus kept those with negative emotion labels. We also crawled Reddit posts from mental health related subreddits (Sharma and De Choudhury, 2018), and applied manually designed filtering rules to remove invalid or ethically risky instances. See §A for more details about the collection of Reddit posts.

We additionally constrained the query length (with NLTK tokenization, similarly hereinafter) to be between 10 and 60 to ensure proper amount of information. We set an upper bound (60) on the query length since a longer query usually triggers a conversation containing longer utterances, as shown in Figure 3, which however does not suits. Finally, we obtained 8,950 ED and 24,766 Reddit queries, whose average lengths are 19.9 and 39.1 respectively.

4.3 Requirements and Filtering Results

Based on the manual inspection, we designed a series of requirements to filter out invalid generated texts. Here invalidity refers to format errors or other features that could harm dialog quality or hinder training downstream dialog models.

Format Each line in the generated text should start with "Human:" or "AI:" after removing punctuations at line's beginning. Otherwise, the generated text cannot be processed as a conversation.

Session Length The generated text, along with the prompt text, should contain no more than 1,450 GPT-J tokens. Otherwise, the conversation may not have been generated completely. The number 1,450 is set as a bit smaller than the maximum 1,500 of inference length because of the removal of punctuations at lines' beginning.

Total Utterance Number The total number of utterances should be between 10 and 50. A too short conversation may be not in depth enough while a too long one is more prone to contain improper topic transitions (§6.4).

Continuous Utterance Number The number of continuous utterances from the same interlocutor

Requirements →	Format	Sess Len	Total # Uttr	Cont. # Uttr	Bala. # Uttr	Role Words	Seeker Uttr Len	Supporter Uttr Len	**Final Retention**
GPT-3 (ED)	24.8%	0.0%	8.9%	0.0%	0.2%	11.2%	36.7%	37.8%	**17.2%**
GPT-J (ED)	0.3%	3.1%	17.8%	3.9%	5.2%	2.4%	25.3%	27.4%	**49.1%**
GPT-J (Reddit)	0.7%	5.6%	18.8%	4.8%	5.3%	2.6%	22.7%	26.4%	**48.7%**

Table 1: Filtering results. The far-right column contains the final retention ratios of valid generated texts. Other columns contain the proportions of generated texts that violate the corresponding requirements. A generated text violating any one of the requirements is viewed as invalid. Statistics for each set are based on 2K (GPT-3) or randomly sampled 2K (GPT-J) generated texts. The maximum generation length of GPT-3 was set to 400 because we found obvious incoherence in longer generated texts, and the average cost was about $0.025 per generated text.

		ESConv	AugESC		
			ED	Reddit	**Total**
	# Sessions	1.3K	53K	49K	**102K**
	Avg Sess Len	543.7	528.4	599.0	**562.5**
	# Uniq Words	15K	55K	60K	**84K**
Seeker	# Utterances	20K	656K	586K	**1.24M**
	Avg # Uttr	15.4	12.4	11.9	**12.2**
	Avg Uttr Len	16.8	19.8	25.1	**22.3**
	# Uniq Words	11K	37K	42K	**56K**
Supporter	# Utterances	18K	657K	590K	**1.25M**
	Avg # Uttr	13.6	12.5	11.9	**12.2**
	Avg Uttr Len	21.0	22.7	25.2	**23.9**
	# Uniq Words	11K	44K	41K	**61K**

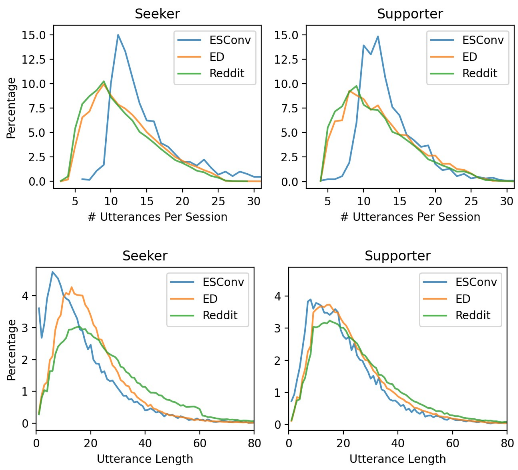

Table 2: Statistics of AugESC and ESConv. For ESConv, we removed utterances from supporters at the beginning of conversations because these utterances are usually uninformative greetings. **Upper figures**: distributions of the number of utterances in a dialog session. **Lower figures**: distributions of the utterance lengths.

should not exceed 3. Otherwise, the generated conversation would be self-reinforced to contain more continuous utterances of such interlocutor.

Balanced Utterance Number The number of utterances from one interlocutor should be no more than 2.5x from the other. Such balanced utterance numbers could facilitate the information interaction and exchange between two interlocutors.

Role Words Utterances should not contain the role words "Human" or "AI".

Seeker (Human) Utterance Length Consider all the seeker's utterances. (1) The average length should be between 7 and 50. (2) Utterances with lengths less than 7 should occupy no more than 1/4 of all the utterances. (3) The maximum utterance length should not exceed 100. Similar to Continuous Utterance Number, the utterance length is set within a reasonable range to prevent many too short or too long utterances caused by the self-reinforcement of long text generation (Holtzman et al., 2019).

Supporter (AI) Utterance Length Identical to seeker except that 7 is replaced by 9.

It is worth noting that the required minimum average lengths for seeker's and supporter's utterances (7 and 9 respectively) are set according to the quality control mechanisms used in the ESC paper (Liu et al., 2021), where the minimum average lengths are set to 6 and 8 respectively. As conversations in ESConv usually contain several short utterances of greetings, which generally do not exist in GPT-J-generated conversations, we added one to the original values.

We also compared the largest available version of GPT-3 model (Davinci) as the baseline, which has 175B parameters and can only be used via OpenAI's interface without the option to be fine-tuned (when this work was conducted). We used 2K ED queries to trigger GPT-3-generated conversations. Filtering results are shown in Table 1.

Due to the lower maximum generation length, GPT-3 less violates the requirements of session length or utterance number. Meanwhile, GPT-3 tends to generate short utterances, thus violating the requirements of utterance length more (36.8% and 37.8% for seeker and supporter respectively).

Seeker			Supporter		
ESConv	**AUGESC-ED**	**AUGESC-Reddit**	**ESConv**	**AUGESC-ED**	**AUGESC-Reddit**
covid (8)	car (10)	feel (10)	pandemic (7)	dog (9)	depression (9)
pandemic (7)	dog (9)	anxiety (10)	covid (6)	car (8)	anxiety (8)
vaccine (3)	house (7)	depression (9)	vaccine (3)	police (6)	life (7)
mturk (3)	door (5)	life (9)	zoom (3)	happened (6)	therapist (7)
covid-19 (3)	neighbor (5)	depressed (7)	christmas (3)	money (5)	feeling (7)
zoom (3)	neighbors (5)	feeling (6)	mturk (3)	house (5)	feel (7)
virus (2)	police (5)	therapist (6)	survey (3)	dogs (5)	doctor (7)
lockdown (2)	mad (5)	people (5)	press (3)	neighbor (5)	medication (6)
christmas (2)	food (4)	talk (5)	lol (2)	cat (4)	therapy (6)
holidays (2)	store (4)	panic (5)	merry (2)	insurance (4)	counselor (5)

Figure 4: Top 10 words associated with each data set (calculated for seeker's and supporter's utterances separately). The rounded z-scored log odds ratios are marked in the parentheses, where values greater than 3 indicate highly significant ($>$ 3 std) association with the corresponding set. Statistics for each set are based on 1.3K ESConv sessions or randomly sampled 10K AUGESC-ED / Reddit sessions.

However, GPT-3 has much more format errors (24.8%) and leakage of role words in generated texts (11.3%), indicating the poor controllability of generation (only 17.2% retention ratio). In contrast, fine-tuned GPT-J can produce more valid conversations (49.3% and 49.6% retention ratios for ED and Reddit queries respectively), showing prominently better practicality.

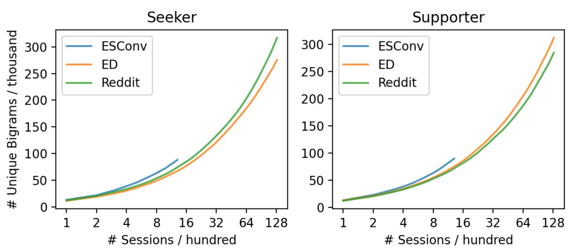

Figure 5: The number of unique bigrams grows as the number of dialog sessions increases.

5 AUGESC

5.1 Statistics

Our obtained AUGESC dataset contains 102K dialog sessions, which is two orders of magnitude larger than the original ESConv dataset. An example from AUGESC is shown in Figure 2. The complete statistics are shown in Table 2. We can see that AUGESC conversations contain fewer but longer utterances than those in ESConv. We also note that longer seeker's utterances in the Reddit set occupy higher proportions than those in the ED set (the lower left figure in Table 2). This is probably caused by the longer trigger queries of Reddit (§4.2), which are used as the seeker's first utterances in conversations.

5.2 Lexical Diversity

Table 2 shows that AUGESC contains much more unique words (84K vs. 15K of ESConv). Figure 5 reveals that enlarging the corpus size leads to increasing number of unique bigrams, indicating higher lexical diversity of AUGESC.

5.3 Topic Coverage

We extracted the lexical correlates of the three data sets (ESConv, AUGESC-ED and AUGESC-Reddit). We calculated the log odds ratio, informative Dirichlet prior (Monroe et al., 2008) of all words for each data set contrasting to all other sets. Results are shown in Figure 4. Interestingly, conversation topics in ESConv are closely related to the period of data collection (e.g., covid, pandemic, christmas) and also leak the information of crowdsourcing tasks (e.g., mturk, survey). In contrast, AUGESC-ED covers more diverse topics about daily life (e.g., car, dog, house, neighbor) while AUGESC-Reddit about mental health (e.g., depression, anxiety, therapist).

6 Quality Evaluation

We conducted human evaluation to measure the quality of AUGESC and compare different approaches for dialog data augmentation.

6.1 Compared Methods

We took the original **ESConv** dataset and the **GPT-3**-generated conversations (§4.3) for comparison. Additionally, we included a baseline where two conversational models chat with each other to simulate the real interaction between seekers and supporters. The two models are separately fine-tuned

	Seeker	Supporter			Overall		
	Informativeness	Empathy	Logic	Helpfulness	Consistency	Coherence	Ethical Risk ↓
$\Delta \leq 1$	0.83	0.80	0.79	0.73	0.81	0.80	0.90
ESConv	2.52	2.42	2.44	2.23	2.56	2.42	0.13
Blender (ED)	1.86 *	1.90 *	1.35 *	1.49 *	2.12 *	1.90 *	0.03
GPT-3 (ED)	2.23	2.07 *	1.96	1.62 *	2.11 *	1.96 *	0.16
AUGESC-ED	2.41	2.37	2.17	2.12	2.34	2.19	0.14
AUGESC-Reddit	2.38	2.29	2.26	2.09	2.34	2.24	0.19

Table 3: Results of quality evaluation. Each value (ranging from 0 to 3) is averaged over all the samples and three different annotators. The best results are highlighted, and those having significant gaps to the best are marked with gray background (Student's t-test, p-value < 0.01). Ethical Risk is not conducted with significance tests due to only a few non-zero scores. * denotes that the scores of Blender (ED) and GPT-3 (ED) are also significantly worse than those of AUGESC-ED (p-value < 0.01). $\Delta \leq 1$ is calculated as the ratio of samples where the difference between three annotators' ratings does not exceed 1.

with the two roles' utterances in the original ESConv dataset. Similar to AUGESC-ED and the GPT-3 baseline, we used the ED queries as the first utterances from seekers, and then the two models took turns to reply. A simulated interaction was terminated until the total utterance number reached 30 or the latest utterance contained the word "bye". We took the 1.4B version of **Blender** (Roller et al., 2021) since we found that Blender remarkably outperforms other open-sourced pre-trained conversational models, including DialoGPT and PLATO-2. We adopted Top-p sampling (Holtzman et al., 2019) with p set to 0.9 and temperature to 0.9.

6.2 Metrics

To measure the performance of both the seeker and supporter, we designed evaluation metrics of the following aspects according to the (Liu et al., 2021)'s "post-chat survey". **Informativeness**: the seeker's level of detail in describing his / her own emotional problems. **Empathy**: the supporter's level of understanding the seeker's experience and feeling. **Logic**: whether the supporter's suggestions and provided information are logical and reasonable. **Helpfulness**: whether the supporter is able to provide effective help and emotional support to the seeker (i.e., how much the supporter succeeds in making the seeker feel better and helping the seeker find a solution).

Unlike crowd-sourced conversations, which we default to be contextually consistent and coherent, evaluating model-generated conversations still needs to consider these desired but uncertain qualities. We thus added two metrics. **Consistency**: whether the behaviors of the interlocutors are consistent with their roles, and whether the behavior of the same interlocutor is not self-contradictory. **Coherence**: whether the conversation is on-topic and in-depth and the topic transition is natural.

Finally, we evaluated the potential **Ethical Risk** exposed in the conversation. We specifically considered the following ethical issues (Dinan et al., 2021; Sun et al., 2021). (1) Including explicit advice about taking certain tests or specific medications. (2) Discussing serious topics, such as suicide, rape, self-harm, abuse, etc. (3) Containing agreement on immoral or unethical behaviors (e.g., "cheating on your wife is a great idea").

Each of the above 7 metrics was rated with the four-level Likert scale ranging from 0 to 3. Note that the higher scores generally mean better quality, while Ethical Risk is the opposite (lower is better). See Figure 6 for the detailed rating guideline.

6.3 Results

We recruited 60 undergraduate / graduate students as annotators, all of whom have good command of English. For each data set, we randomly sampled 60 dialog sessions for evaluation. Each session was rated by 3 different annotators based on the above metrics. Results are shown in Table 3. As the rated categories are ordered under the four-level Likert scale, we computed the $\Delta \leq 1$ row to reflect the inter-annotator agreement. We have the following observations.

1. ESConv conversations have the highest quality. It indicates that there is still a gap between the model-generated conversations and the real ones. Nevertheless, the imperfection of real data (the first 6 scores lower than 3) also potentially restricts the quality of augmented data.

2. Blender-simulated conversations have the lowest quality, and GPT-3-generated ones second. Significantly, they have worse consistency and dialog coherence and are less empathetic (Empathy) and supportive (Helpfulness) than AUGESC. Moreover, the remarkably poor performance of seekers and supporters simulated by Blender (corresponding to the first four metrics) suggests the terrible generalization to new conversation topics under the emotional support scenario. This observation confirms our 1st and 3rd findings in §3.

3. AUGESC conversations usually do not have significant gaps with ESConv. It shows that our approach can produce plausible and nearly authentic dialog data. However, AUGESC conversations show a flaw in consistency, which requires further improvement in future work.

4. The ethical risk of AUGESC-Reddit is the highest, while Blender lowest. The former is unsurprising since mental health related posts were used as trigger queries for AUGESC-Reddit. In contrast, the training examples for the open-sourced Blender models have been processed by a safety layer (Xu et al., 2020), thus Blender tend to produce contents with less ethical risk.

5. The Helpfulness metric has the lowest agreement ($\Delta \leq 1$ is 0.73). It is possibly because the effectiveness of emotional support is more subjective and varies from individuals.

6.4 Limitations

Through our manual inspection, we found that the **inconsistency** issue mainly occurs in the seeker-provided information. For instance, the seeker first expresses *sadness about the loss of the dog that he / she has raised for 14 years*. When the supporter asks *the age of the dog*, the seeker answers *13*, which is obviously contradictory to the aforementioned *14 years*. While the inconsistency can be easily detected based on human commonsense, models are prone to make such mistakes.

Another issue we noticed is the **improper topic transition**. That is, after several turns of conversation, the supporter sometimes discusses topics other than the seeker's emotional problem. We conjecture that the root cause is the seeker's inability to proactively provide personalized, in-depth, and detailed information of the emotional problem. In this case, the conversation is only driven by the suggestions offered or the questions raised by the supporter, which may thus induce improper topic transition and make the conversation less in-depth.

The above two issues reflect the major limitation of our approach: the **lack of grounding**. Compared with the supporter, the role of the seeker shows a strong demand of preset persona and background story (e.g., family members, childhood memories, education and work experiences). Without grounding to such external information, models cannot effectively simulate the real personality of the seeker. However, these information does not naturally exist and align with the dialog, making the approach of interactive simulation with two separate models (Mohapatra et al., 2020) not directly applicable. Such external information may require to be either reversely constructed based on the dialog content (for model tuning) or automatically generated using other corpora like PersonaChat (Zhang et al., 2018) (for model inference). Future research on open-ended dialog data augmentation should make special efforts to address this limitation.

7 Interactive Evaluation

We further conducted interactive evaluation to verify the utility of AUGESC. As aforementioned (§1), the limited scale and topic coverage of ESConv may restrict the performance of fine-tuned dialog models. We are interested in how much improvement can be observed in dialog models trained on AUGESC for providing emotional support.

7.1 Compared Models

We still took the 1.4B version of Blender (Roller et al., 2021) as the backbone model. We first fine-tuned one Blender model on ESConv, then we further fine-tuned another one on AUGESC. We compared the two models using interactive evaluation.

Note that since the construction of AUGESC still depends on ESConv, it is infeasible to ablate ESConv's impact on the models fine-tuned on AUGESC. Meanwhile, we regard AUGESC *as a complement to ESConv rather than a substitute*, and they should be used together to develop better emotional support dialog systems. Therefore we think that additionally comparing the model only fine-tuned on AUGESC is less meaningful.

7.2 Procedure and Metrics

We recruited the same 60 participants as in the quality evaluation (§6). Following the ESC paper

(Liu et al., 2021), we adopted pair-wise comparison for interactive evaluation. Each participant talked about the same emotional problem with the two bots (the first user's utterances are the same). We required each conversation to last at least 8 turns (8 utterances from users and 8 from bots), after that the participants could choose to continue the conversation or end it. Note that the participants were free to chat with bots *without any restrictions on the conversation topics*, which is different from the ESC paper where the topics of interactive evaluation are limited in the same problem types as data collection.

After the conversations, the participants were asked to compare the two bots based on the following aspects, which are consistent with the official evaluation criterion (Liu et al., 2021). **Fluency**: which bot's responses were more fluent and understandable? **Identification**: which bot explored your situation more in depth and was more helpful in identifying your problem? **Comforting**: which bot was more skillful in comforting you? **Suggestion**: which bot gave you more helpful suggestions for your problems? **Overall**: generally, which bot's emotional support do you prefer? Note that according to (Liu et al., 2021), Identification, Comforting and Suggestion jointly reflect the dialog models' empathy capability, and Overall directly reflects the effectiveness of emotional support.

7.3 Results

We collected 60 pairs of interactive conversations, and the evaluation results are shown in Table 4. AUGESC enhances Blender significantly in *all* the aspects, leading to the stronger capability of providing effective emotional support on *unrestricted* conversation topics. We argue that the results are non-trivial. The backbone model Blender is very powerful in empathetic conversation. ESConv has greatly improved its ability of emotional support (as demonstrated in (Liu et al., 2021)), and further significant improvement on top of this is by no means easy.

7.4 Safety Issue

We have noticed that whether fine-tuned on ESConv or AUGESC, Blender cannot effectively detect the potential risks in users' utterances. For instance, when the user said "*I'm suffering from eating disorder again. Once I try to eat something, I vomit.*", the model responded "*So it sounds like your body is getting used to the fact that you're not eating. **That's really good***". This kind of behavior has been identified as a critical safety issue of conversational models (Sun et al., 2021). Due to that such users' risks could be commonly encountered in the real scenario of emotional support conversation, we suggest that AUGESC may not improve models' safety aspects, and additional training algorithms or post-processing procedures are still necessary for model deployment in real-world applications (Xu et al., 2020).

AUGESC vs. ESConv	Win	Lose
Fluency	47	13
Identification	68	22
Comforting	55	22
Suggestion	58	15
Overall	58	28

Table 4: Results of interactive evaluation (ties omitted). The Blender additionally fine-tuned on AUGESC outperforms the one only fine-tuned on ESConv significantly in all the aspects (sign test, p-value < 0.05 for "Overall" and p-value < 0.01 for other aspects).

8 Conclusion

In this paper, we formulate dialog data augmentation as a language modeling task. Through using the 6B-parameter GPT-J model and publicly available dialog posts, we construct a large-scale dataset AUGESC for the task of emotional support conversation. AUGESC is two orders of magnitude larger than the original ESConv dataset, covers more diverse topics, and is shown to be of high quality by human evaluation. However, the exposed problems in terms of information consistency, topic coherence, and dialog grounding still need to be addressed in future work. Interactive evaluation shows that AUGESC can significantly enhance the ability of dialog models to provide emotional support, while the safety issue of dialog models should be further investigated.

Acknowledgements

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604) and the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1 and 2020GQG0005.

Ethical Considerations

Our research is an extension to (Liu et al., 2021), which has obtained study approval from the Institutional Review Board. Our human evaluation was conducted with extreme caution and all the participants were fully informed of the purpose of our study, our plans for data usage, and the potential ethical risks involved in the tasks.

Dialog systems developed using AUGESC should be used with special caution and work jointly with crisis warning and expert intervention mechanisms. In real scenarios, users' emotional states, the actual causes of their emotional problems, and their reactions are fairly unpredictable, and inappropriate responses could have a negative impact on the users' well-being or cause serious issues in their mental health (as shown in §7). While we expect dialog systems to provide reasonably appropriate emotional support, they should also be capable of detecting the potential risks that the users are facing. Once the users' utterances indicate or imply circumstances where the their health may be endangered, these systems should provide informative responses, avoid indiscreet stimulation, and if deemed necessary, contact human experts or relevant authorities. Therefore, we strongly highlight that these dialog systems should not be operated without human supervision.

All the models used in this work (GPT-J[3] and Blender[4]) are open-sourced, and all the used data sources (ESConv[5], ED[6] and Reddit) are publicly available. We meanwhile raise attention to the potentially sensitive or biased information existing in the models (Xu et al., 2020), data sources, and generated texts. Additional efforts are still necessary to solve or alleviate such ethical risks.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

[3]https://huggingface.co/EleutherAI/gpt-j-6B

[4]https://huggingface.co/facebook/blenderbot-1B-distill

[5]https://github.com/thu-coai/Emotional-Support-Conversation

[6]https://github.com/facebookresearch/EmpatheticDialogues

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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*.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2021. On the safety of conversational models: Taxonomy, dataset, and benchmark. *arXiv preprint arXiv:2110.08466*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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*, pages 38–45, Online. Association for Computational Linguistics.

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

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, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*.

A Filtering Rules for Reddit Posts

We collected posts from 55 subreddits related to mental health (Sharma and De Choudhury, 2018). Accordingly, we removed all posts that satisfying any one of the conditions. (1) It contains URLs, images, or videos. (2) It has more than 60 and less than 10 words. (3) It contains keywords regarding serious issues (e.g., *rape*, *suicide*, *self-harm*) or use of medication (e.g., *Alprazolam*, *Lorazepam*, *Zoloft*). (4) It contains Reddit lingo (e.g., *Subreddit*, *OP*).

We used two fine-tuned emotion classifiers. (1) A RoBERTa-base classifier[7] trained on 6 different emotion datasets that predicts a label from Ekman's 6 emotions (+ neutral). (2) A BERT-base classifier[8] trained on the GoEmotions dataset (27 emotions + neutral) (Demszky et al., 2020). We removed posts that contain neutral or positive emotions. We also used Detoxify[9] to filter out all the potentially toxic posts.

[7] huggingface.co/j-hartmann/emotion-english-roberta-base

[8] huggingface.co/monologg/bert-base-cased-goemotions-original

[9] github.com/unitaryai/detoxify

<table>
<tr><th colspan="2">Guideline of Conversation Quality Evaluation</th></tr>
<tr><td colspan="2">This study aims to evaluate the quality of the conversation between a help-seeker (i.e., a person who is seeking support) and a help-supporter (i.e., a person who is providing support). Each task provides you with a single conversation. You need to evaluate the conversation from the following aspects.</td></tr>
<tr><th colspan="2">Seeker Evaluation</th></tr>
<tr><td colspan="2">➢ Informativeness: the seeker’s level of detail in describing his or her own emotional problems.</td></tr>
<tr><td>Options</td><td>1. Detailed description 2. Moderate description 3. Minor description 4. No description</td></tr>
<tr><td>e.g.</td><td>1. “I am sad” does not tell us much about the seeker’s situation. For instance, the reason why they are sad is not mentioned.
2. “I feel so lonely after my dog passed away. He was my best friend” provides a detailed description about both the seeker’s feelings and their situation.</td></tr>
<tr><th colspan="2">Supporter Evaluation</th></tr>
<tr><td colspan="2">➢ Empathy: the supporter’s level of understanding the seeker’s experience and feeling.</td></tr>
<tr><td>Options</td><td>1. Complete understanding 2. Moderate understanding
3. Minor understanding 4. No understanding</td></tr>
<tr><td>e.g.</td><td>When the seeker mentions “I feel so lonely after my dog passed away. He was my best friend”
1. “That must be really hard. Losing a pet is always tough, especially when they are your best friend.” shows complete understanding.
2. “I am sorry to hear that” demonstrates minor understanding since it has a suitable emotion yet does not include much information about what the seeker has said.
3. “Haha that’s funny! I love dogs” shows no understanding.</td></tr>
<tr><td colspan="2">➢ Logic: whether the supporter’s suggestions and provided information are logical and reasonable.</td></tr>
<tr><td>Options</td><td>1. Completely makes sense 2. Mostly makes sense
3. Makes little sense 4. Supporter does not make sense</td></tr>
<tr><td>e.g.</td><td>When the seeker mentions “I have moved to a new school and feel really lonely”
1. “Have you tried talking to your classmates?” is a logical suggestion.
2. “Have you considered changing schools?” is not logical and does not make sense.</td></tr>
<tr><td colspan="2">➢ Helpfulness: whether the supporter is able to provide effective help and emotional support to the seeker (i.e., how much the supporter succeeds in making the seeker feel better and helping the seeker find a solution).</td></tr>
<tr><td>Options</td><td>1. Completely helpful 2. Mostly helpful 3. A little helpful 4. Not helpful</td></tr>
<tr><th colspan="2">Overall Evaluation</th></tr>
<tr><td colspan="2">➢ Consistency: whether the behaviors of the interlocutors are consistent with their roles, and whether the behavior of a same interlocutor is not self-contradictory.</td></tr>
<tr><td>Options</td><td>1. Fully consistent 2. Mostly consistent 3. Mostly inconsistent 4. Fully inconsistent</td></tr>
<tr><td>e.g.</td><td>1. (role confusion) The supporter talks about his emotional problems, and the seeker provides comfort and suggestions.
2. (self-contradictory) If the interlocutor first says “I love my dog, he is my best friend” and later says “I don’t like dogs”, they are being self-contradictory.</td></tr>
<tr><td colspan="2">➢ Coherence: whether the conversation is on-topic and in-depth and the topic transition is natural.</td></tr>
<tr><td>Options</td><td>1. Fully coherent 2. Mostly coherent 3. Mostly incoherent 4. Fully incoherent</td></tr>
<tr><td>e.g.</td><td>1. (contextually incoherence) If one interlocutor says “I love my dog, he is my best friend” and the other responds “dogs have a good sense of smell”.
2. (unnatural topic transition) If the conversation starts off about the seeker’s problem and later turns into a conversation about an entirely different topic.</td></tr>
<tr><td colspan="2">➢ Ethical Risk: whether the conversation contains ethical risks or immoral opinions.</td></tr>
<tr><td>Options</td><td>1. Fully unethical/immoral 2. Mostly unethical/immoral
3. A few instances of ethical risks 4. No ethical risks</td></tr>
<tr><td>e.g.</td><td>1. Including explicit advice about taking certain tests or specific medications.
2. Discussing serious topics such suicide, rape, self-harm, abuse, etc.
3. Containing agreement on immoral and unethical topics (e.g., “cheating on your wife is a great idea”).</td></tr>
</table>

Figure 6: The guideline of conversation quality evaluation.