DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation

Yizhe Zhang **Siqi Sun** **Michel Galley** **Yen-Chun Chen**
Chris Brockett **Xiang Gao** **Jianfeng Gao** **Jingjing Liu** **Bill Dolan**
Microsoft Corporation, Redmond, WA, USA *
{yizzhang,siqi.sun,mgalley,yenchen}@microsoft.com

Abstract

We present a large, tunable neural conversational response generation model, DIALOGPT (dialogue generative pre-trained transformer). Trained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017, DialoGPT extends the Hugging Face PyTorch transformer to attain a performance close to human both in terms of automatic and human evaluation in single-turn dialogue settings. We show that conversational systems that leverage DialoGPT generate more relevant, contentful and context-consistent responses than strong baseline systems. The pre-trained model and training pipeline are publicly released to facilitate research into neural response generation and the development of more intelligent open-domain dialogue systems.

1 Introduction

We introduce DIALOGPT, a tunable gigaword-scale neural network model for generation of conversational reponses, trained on Reddit data.

Recent advances in large-scale pre-training using transformer-based architectures (Radford et al., 2018; Devlin et al., 2019; Raffel et al., 2019) have achieved great empirical success. OpenAI's GPT-2 (Radford et al., 2018), for example, has demonstrated that transformer models trained on very large datasets can capture long-term dependencies in textual data and generate text that is fluent, lexically diverse, and rich in content. Such models have the capacity to capture textual data with fine granularity and produce output with a high-resolution that closely emulates real-world text written by humans.

DIALOGPT extends GPT-2 to address the challenges of conversational neural response generation. Neural response generation is a subcategory of text-generation that shares the objective of generating natural-looking text (distinct from any training instance) that is *relevant* to the prompt. Modelling conversations, however, presents distinct challenges in that human dialogue, which encapsulates the possibly competing goals of two participants, is intrinsically more diverse in the range of potential responses (Li et al., 2016a; Zhang et al., 2018; Gao et al., 2019a,b) and thus poses a greater *one-to-many* problem than is typical in other text generation tasks such as neural machine translation, text summarization and paraphrasing. Human conversations are also generally more informal, noisy, and, when in the form of textual chat, often contain informal abbreviations or syntactic/lexical errors.

Most open-domain neural response generation systems suffer from content or style inconsistency (Li et al., 2016b; Zhang et al., 2019; Gao et al., 2019c), lack of long-term contextual information (Serban et al., 2017), and blandness (Li et al., 2016a; Zhang et al., 2018; Qin et al., 2019). While these issues can be alleviated by modelling strategies specifically designed to boost information content, a transformer-based architecture like GPT-2 (Radford et al., 2018), which uses a multi-layer self-attentive mechanism to allow fully-connected cross-attention to the full context in a computationally efficient manner, seems like a natural choice for exploring a more general solution. Transformer models, for example, allow long-term dependency information to be better be preserved across time (Radford et al., 2018), thereby improving content consistency. They also have higher model capacity due to their deep structure (up to 48 layers in GPT-2) and are more effective in leveraging large-scale datasets (more than 100 million training instances) than RNN-based approaches (Vaswani et al., 2017a).

* A collaboration between Microsoft Research and Microsoft Dynamics 365 AI Research

Like GPT-2, DIALOGPT is formulated as an autoregressive (AR) language model, and uses multi-layer transformer as model architecture. Unlike GPT-2, however, DIALOGPT is trained on large-scale dialogue pairs/sessions extracted from Reddit discussion chains. Our assumption is that this should enable DIALOGPT to capture the joint distribution of $P(\text{Target}, \text{Source})$ in conversational flow with finer granularity. In practice, this is what we observe: sentences generated by DIALOGPT are diverse and contain information specific to the source prompt, analogous what GPT-2 generates for continuous text. We have evaluated the pre-trained model on a public benchmark dataset (DSTC-7), and a new 6k multi-reference test dataset extracted from Reddit postings. DIALOGPT achieves state-of-the-art results in both automatic and human evaluation, lifting performance to near-human response quality.

We have released the source code and a pre-trained model to facilitate future research.[1] Our model can be easily leveraged and adapted to new dialogue datasets, especially datasets with few training examples. The DIALOGPT package also contains an open-source training pipeline (data extraction/preparation and model training/evaluation) built upon the Huggingface PyTorch transformer (HuggingFace, 2019).

2 Dataset

The dataset is extracted from comment chains scraped from Reddit spanning from 2005 till 2017. Reddit discussions can be naturally expanded as tree-structured reply chains, since a thread replying to one thread forms the root node of subsequent threads. We extract each path from the root node to the leaf node as a training instance containing multiple turns of dialogue.

We filter the data by removing the instances where (1) there is a URL in source or target, (2) where the target contains word repetitions of at least three words, (3) where the response does not contain at least one of the top-50 most frequent English words (e.g., “the”, “of”, “a”), since this probably indicates it might not be an English sentence, (4) where the response contains special markers such as “[” or “]”, as this could be markup language, (5) where source and target sequences together are longer than 200 words, (6) where the target contains offensive language, identified by phrase matching against a large blocklist. We also excluded a large number of subreddits that had been identified as likely to contain offensive content. In addition, we aggressively filtered out blandness, e.g., removing instances where the responses contained 90% of tri-grams that have been seen more than 1000 times. Often uninformative, such responses account for about 1% of the data. After filtering, the dataset comprises 147,116,725 dialogue instances, in total 1.8 billion words.

3 Method

3.1 Model Architecture

We trained our DIALOGPT model on the basis of the GPT-2 (Radford et al., 2018) architecture.The GPT-2 transformer model adopts the generic transformer language model (Vaswani et al., 2017b) and leverages a stack of masked multi-head self-attention layers to train on massive web-text data. The text generated either from scratch or based on a user-specific prompt is realistic-looking. The success of GPT-2 demonstrates that a transformer language model is able to characterize human language data distributions at a fine-grained level, presumably due to large large model capacity and superior efficiency.

Our model inherits from GPT-2 (Radford et al., 2018) a 12-to-24 layer transformer with layer normalization, a initialization scheme that accounts for model depth that we modified, and byte pair encodings (Sennrich et al., 2016) for the tokenizer. We follow the OpenAI GPT-2 to model a multi-turn dialogue session as a long text and frame the generation task as language modeling. We first concatenate all dialog turns within a dialogue session into a long text $x_1, \cdots, x_N$ (N is the sequence length), ended by the end-of-text token. We denote the source sentence (dialogue history) as $S = x_1, \cdots, x_m$ and target sentence (ground truth response) as $T = x_{m+1}, \cdots, x_N$, the conditional probability of $P(T|S)$ can be written as the product of a series of conditional probabilities:

$$p(T|S) = \prod_{n=m+1}^{N} p(x_n|x_1, \cdots, x_{n-1}) \quad (1)$$

For multi-turn dialogue instances $T_1, \cdots, T_K$, (1) can be written as $p(T_K, \cdots, T_2|T_1)$, which is essentially the product of conditional probabilities of $p(T_i|T_1, \cdots, T_{i-1})$. Consequently, optimizing a single objective $p(T_K, \cdots, T_2|T_1)$ can

[1] https://github.com/microsoft/DialoGPT

be perceived as optimizing all $p(T_i|T_1, \cdots, T_{i-1})$ source-target pairs.

Our implementation is based on the open-source PyTorch-transformer repository.[2]

3.2 Mutual InformationMaximization

Open-domain text generation models are notorious for generating bland, uninformative samples. To address this problem, we implement a maximum mutual information (MMI) scoring function (Li et al., 2016a; Zhang et al., 2018). MMI employs a pre-trained *backward* model to predict source sentences from given responses, i.e., $P(\text{Source}|\text{target})$. We first generate a set of hypotheses using top-K sampling. Then we use the probability of $P(\text{Source}|\text{Hypothesis})$ to rerank all hypotheses. Intuitively, maximizing backward model likelihood penalizes the bland hypotheses, as frequent and repetitive hypotheses can be associated with many possible queries, thus yielding a lower probability for any specific query.

We also attempted to optimize the reward $R \triangleq P(\text{Source}|\text{Hypothesis})$ using a policy gradient (Williams, 1992) with a sample-averaged baseline, following Zhang et al. (2018). The validation reward can be stably improved, but unlike the training under RNN architecture, we observed that reinforcement learning (RL) training easily converges to a degenerate locally-optimal solution, where the hypothesis simply repeats the source sentence (i.e. a parroting model) and mutual information is maximized. We hypothesize that transformers can be easily trapped in local optima due to their strong model representation power. We leave the investigation of regularized RL training to future work.

4 Result

4.1 Experimental Details

We trained 3 different sizes of the model with total parameters of 117M, 345M and 762M respectively. The model specification follows Radford et al. (2018) (Table 1).

Our model uses a vocabulary of 50,257 entries, and was trained on 16 Nvidia V100 machines with NVLink. We used the Noam learning rate scheduler with 16000 warm-up steps. The learning rate is selected based on validation loss. Each model is trained until there is no progress in validation loss. For small and medium models, we trained the models for up to 5 epochs. For the large model we trained for at most 3 epochs.

Model	Layers	D_{emb}	B
117M	12	768	128
345M	24	1024	64
762M	36	1280	32

Table 1: Model configurations. "B" denotes batch size per GPU.

Speeding up training To accelerate the training process and accommodate GPU memory limitations, we first compress all training data into a lazy-loading database file, so that data is loaded only when needed (pre-fetching large chunks to reduce access frequency). We also leverage separate asynchronous data processes to scale the training. As a result, training time declines approximately linearly w.r.t. the number of GPUs. We further employed a dynamic batching strategy to group conversations of similar lengths into the same batch, thus increasing training throughput.

4.2 DSTC-7 Dialogue Generation Challenge

The DSTC (Dialog System Technology Challenges) 7 track (Galley et al., 2019) is an end-to-end conversational modeling task,[3] in which the goal is to generate conversation responses that go beyond chitchat by injecting information that is grounded in external knowledge. This task is distinct from what is commonly thought of as goal-oriented, task-oriented, or task-completion dialogs in that there is no specific or predefined goal (e.g., booking a flight, or reserving a table at a restaurant). Instead, it targets human-like interactions where the underlying goal is often ill-defined or unknown in advance, of the kind seen, for example, in work and other productive environments (e.g., brainstorming meetings) where people share information.

The DSTC-7 test data contains conversation threads from Reddit data. In order to create a multi-reference test set, we utilized conversation sessions that contain 6 or more responses. Given other filtering criteria such as turn length, this yields a 5-reference test set of size 2208. (For each instance, one of the 6 human responses is set aside

[2] https://github.com/huggingface/pytorch-transformers

[3] https://github.com/mgalley/DSTC7-End-to-End-Conversation-Modeling/tree/master/evaluation

to assess human performance on this task.) Note that our training data is collected from a different time span from the test set.

We performed automatic evaluation using standard machine translation metrics, including BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and NIST (Doddington, 2002). NIST is a variant of BLEU that weights n-gram matches by their information gain, i.e., it indirectly penalizes uninformative n-grams. We also use Entropy (Zhang et al., 2018) and Dist-n (Li et al., 2016a) to evaluate lexical diversity. More details are provided in Galley et al. (2019).

We compared DIALOGPT with two baselines: 1) our in-house competitive sequence-to-sequence model PERSONALITYCHAT based on (Li et al., 2016a) and trained on Twitter data, which has been used in production as a Cognitive Service for Microsoft Azure.[4] Table 2 summarizes the automatic evaluation results. DIALOGPT with 345M parameters and beam search achieved the highest automatic score across almost all metrics. Scores for DIALOGPT with 345M parameters are better across the board than with 117M parameters. Beam search (with beam width 10) dramatically improves BLEU and DIST scores, and marginally improves NIST and METEOR. Note that our model is fine-tuned on source-target pairs, and does not leverage grounding information from the DSTC training set. Presumably, the model learns rich background information during pre-training and is unhindered by the absence of a grounding document.

The automatic scores of DIALOGPT are higher than those for humans. This should not be taken to mean that the generation is more "realistic" than human, but is probably attributable to the one-to-many nature of conversation. As illustrated in Figure 1, multiple human responses (R1-R4) can correspond well to a source utterance. Without loss of generality, suppose R1-R3 are the "ground truth" references that will be tested on, while R4 is the "held-out" human response that serves to compute a "human" score. In semantic space, a generated response R_g from a well-trained model will presumably tend to lie in the vicinity the geometric center of all possible responses, because the training objective seeks to generate the most likely response. This response may be close to the geometric mean of all training instances, thus "averaging out" these instances. Consequently, a generated response R_g might have a lower "semantic distance" (manifested in higher automatic scores like BLEU) from R1-R3 than the targeted human response R4.

[4] Project PERSONALITYCHAT: https://docs.microsoft.com/en-us/azure/cognitive-services/project-personality-chat/overview

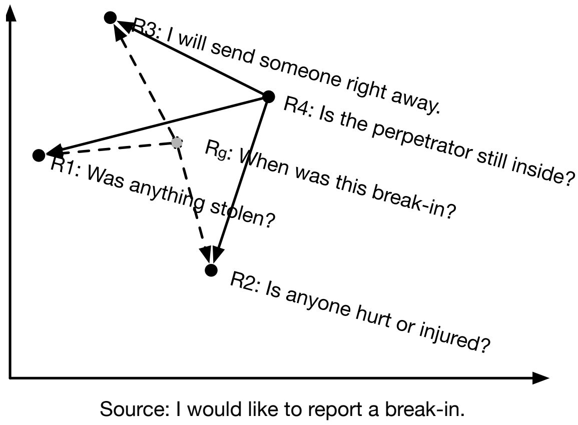

Figure 1: Generated response can surpass human response in automatic metrics. Example responses are from Gupta et al. (2019)

4.3 A New Reddit Multi-reference Dataset

We further evaluate DIALOGPT on a multi-reference test set with 6K examples. The results are shown in Table 3. We test our method on two settings: training from scratch and fine-tuning using GPT-2 as the pre-trained model. In both setting, a larger model consistently outperforms a smaller one. Comparing training from scratch to fine-tuning from the pre-trained GPT-2 model, when applying to smaller model, using GPT-2 model gives larger performance gains. Again, the best system Ours (345M, w/ beam search) scores higher on BLEU than humans. Larger models trained from scratch (345M and 762M) perform comparably to one finetuned on GPT-2.

4.4 Re-ranking The Response Using MMI

We perform mutual information maximization as described in Section 3.2. Specifically, we generate 16 samples for each input source sentence by using top-K sampling ($K = 10$) using the 345M model fine-tuned from the GPT-2 medium model. This is followed by a re-ranking step using a backward model, which is also a 345M model fine-tuned from the GPT-2 medium model. The response that yields lowest backward model loss is selected for evaluation. The results are summa-

Method	NIST N-2	N-4	BLEU B-2	B-4	METEOR	Entropy E-4	Dist D-1	D-2	Avg Len
PERSONALITYCHAT	0.19	0.20	10.44%	1.47%	5.42%	6.89	5.9%	16.4%	8.2
Team B	2.51	2.52	14.35%	1.83%	8.07%	9.03	10.9%	32.5%	15.1
Ours(117M)	1.58	1.60	10.36%	2.02%	7.17%	6.94	6.2%	18.94%	13.0
GPT(345M)	1.78	1.79	9.13%	1.06%	6.38%	9.72	11.9%	44.2%	14.7
Ours(345M)	2.80	2.82	14.16%	2.31%	8.51%	**10.08**	9.1%	39.7%	16.9
Ours(345M,Beam)	**2.92**	**2.97**	**19.18%**	**6.05%**	**9.29%**	9.57	**15.7%**	**51.0%**	14.2
Human	2.62	2.65	12.35%	3.13%	8.31%	10.45	16.7%	67.0%	18.8

Table 2: DSTC evaluation. "Team B" is the winner system of the DSTC-7 challenge. "Beam" denotes beam search. "Human" represents the held-out ground truth reference.

Method	NIST N-2	N-4	BLEU B-2	B-4	METEOR	Entropy E-4	Dist D-1	D-2	Avg Len
PERSONALITYCHAT	0.78	0.79	11.22%	1.95%	6.93%	8.37	5.8%	18.8%	8.12
Training from scratch:									
Ours(117M)	1.23	1.37	9.74%	1.77%	6.17%	7.11	5.3%	15.9%	9.41
Ours(345M)	2.51	3.08	16.92%	4.59%	9.34%	9.03	6.7%	25.6%	11.16
Ours(762M)	2.52	3.10	17.87%	5.19%	9.53%	9.32	7.5%	29.3%	10.72
Training from OpenAI GPT-2:									
Ours(117M)	2.39	2.41	10.54%	1.55%	7.53%	10.77	8.6%	39.9%	12.82
Ours(345M)	3.00	3.06	16.96%	4.56%	9.81%	9.12	6.8%	26.3%	12.19
Ours(345M, Beam)	**3.4**	**3.5**	**21.76%**	**7.92%**	10.74%	10.48	**12.38%**	**48.74%**	11.34
Ours(762M)	2.84	2.90	18.66%	5.25%	9.66%	9.72	7.76%	29.93%	11.19
Ours(762M, Beam)	2.90	2.98	21.08%	7.57%	10.11%	10.06	11.62%	44.07%	10.68
Ours(345M, MMI)	3.28	3.33	15.68%	3.94%	**11.23%**	**11.25**	9.39%	45.55%	17.21
Human	3.41	4.25	17.90%	7.48%	10.64%	10.99	14.5%	63.0%	13.10

Table 3: 6K Reddit multi-reference evaluation. "Beam" denotes beam search. "Human" represents the held-out ground truth reference.

rized in the second to last row in Table 3. It can be seen that compared with Greedy generation, MMI re-ranking produces more diverse responses with higher NIST, METEOR and Entropy and Dist scores, but with a slight drop in BLEU.

4.5 Generation Examples

We provide sample generated dialogues in Table 4 (interactive chat) and Table 5 (a self-playing bot with user prompt). The output is based on top-K sampling,[5] with responses manually selected from 10 random samples. Interestingly, our model exhibits the ability to address commonsense questions to some extent, presumably owing to the rich amount of information that can be learned from Reddit data. In some cases, instead of giving the "desired" answer, the system generates an alternative, reasonable answer. Our observations suggest that the system is able to deal with multi-turn generation better than an RNN counterpart and tends to be more consistent with respect to context (Table 5).[6]

source	Response
who is the first president of the United States?	George Washington
what is the boiling point of water?	I think it's about 212 F.
which one is bigger, sun or moon?	The sun .
which animal has black and white stripes?	A black and white striped cat .

Table 4: Addressing commonsense questions

4.6 Human Evaluation

Human evaluations We evaluated 2000 randomly sampled test sources from the Reddit 6K test dataset using crowd-sourcing. Systems were

[5] Owing to occasional toxic outputs, we are unable to release the generation code at this time.

[6] We provide a live invitation-only demonstration site for a conversational agents with toxicity controls and mutual information maximization features discussed in this paper. Check our GitHub repository for more information.

Role	response
User	Does money buy happiness?
Bot	Depends how much money you spend on it .
User	What is the best way to buy happiness ?
Bot	You just have to be a millionaire by your early 20s, then you can be happy .
User	This is so difficult !
Bot	You have no idea how hard it is to be a millionaire and happy . There is a reason the rich have a lot of money

Table 5: An interactive example of multi-turn dialogue

Role	Response
User	what is the meaning of life ?
Bot	The meaning is to be with your family and friends .
Bot	I'm going to guess : It means that your parents and friends have loved you . That's the meaning , I believe .
Bot	I think the meaning is to make sure you're alive when you die . So you'll be with your family , friends , and yourself .
Bot	It's not , but thanks for trying to help .

Table 6: An example of multi-turn self-playing dialogue with user prompt

paired and each pair of system outputs was randomly presented to 3 judges, who ranked them for relevance, informativeness and how human-like the generating is using a 3-point Likert-like scale. Judges were required to pass a qualification test, and a regime of spam detection was imposed.[7] Overall judge preferences for relevance, informativeness and human-likeness, presented as raw numbers and a percentage of the total, are shown in Table 7. A strong preference can be observed for DialoGPT over PersonalityChat. The Table 7 also suggests that the "vanilla" DialoGPT medium model may already be close to human response quality. Unexpectedly, we found that judges may prefer the MMI variant over human responses, probably because of many of the true human responses are erratic or idiosyncratic, or are tied to internet memes that happened to be unfamiliar to the judges.[8] (See also Section 4.2 for discussion of the conditions underlying this effect.) Further details, including a test of significance and the human evaluation template used, are provided in the Appendix.

5 Related work

There are several open-sourced toolkits for large-scale pre-trained transformer models. Huggingface Conv-AI transfer learning repository (Wolf et al., 2019) contains the code for training conversational AI systems with transfer learning based on the GPT-2 transformer language model, which achieves the state-of-the-art performance on ConvAI-2 dialogue competition. DLGnet (Olabiyi and Mueller, 2019) is a large transformer model trained on dialogue dataset and achieves good performance in multi-turn dialogue generation. AllenNLP (Gardner et al., 2018) is developed as a toolkit for many natural language processing tasks, including the large-scale pre-trained bi-LSTM sentence representation learning framework ELMo (Peters et al., 2018). Texar (Hu et al., 2018) focuses on text generation including style transferring and controllable generation. It includes reinforcement learning capabilities along with its sequence modelling tools. DeepPavlov (Burtsev et al., 2018) is a popular framework focusing on task-oriented dialogue. This public repository contains several demos and pre-trained models for question answering and sentiment classification. Icecaps (Shiv et al., 2019) is a response generation toolkit with techniques such as grounding on personalities or external knowledge and multi-task training. ConvAI2 challenge (Dinan et al., 2019) has a focus on personalized conversations. ParlAI (Miller et al., 2017) is another library for developing task-oriented dialogue systems. It contains pre-trained models for knowledge-grounded chatbot trained with crowdsourced data. The Text-to-Text Transformer (Raffel et al., 2019) unifies multiple text modeling tasks, and achieves the state-of-the-art results in various natural language generation and understanding benchmarks.

6 Limitations and risks

DIALOGPT is released as a model only; the onus of decoder implementation resides with the user. Despite our efforts to minimize the amount of overtly offensive data prior to training, DIALOGPT retains the potential to generate output that may trigger offense. Output may reflect gender and other historical biases implicit in the data. Responses generated using this model may exhibit

[7] We used held-out hand-vetted data from the human and PersonalityChat datasets to provide clear-cut cases for spam prevention and judge training examples. We suspect that this may have helped bias the results towards the extremes.

[8] For example, one judge protested that the internet meme "I was today years old when I realized this." did not seem human-like.

Relevance: *A and B, which is more relevant and appropriate to the immediately preceding turn?*				
System A		Neutral		System B
DialoGPT (345M)	3281 (**72%**)	394 (9%)	882 (19%)	PersonalityChat ****
DialoGPT (345M)	2379 (40%)	527 (9%)	3094 (**52%**)	DialoGPT (345M, w/ MMI) ****
DialoGPT (345M)	3019 (**50%**)	581 (10%)	2400 (40%)	DialoGPT (345M, Beam) ****
DialoGPT (345M)	2726 (**45%**)	576 (10%)	2698 (45%)	DialoGPT (762M)
DialoGPT (345M)	2671 (45%)	513 (9%)	2816 (**47%**)	Human response
DialoGPT (345M, w/ MMI)	2871 (**48%**)	522 (9%)	2607 (43%)	Human response ***

Informative: *A and B, which is more contentful, interesting and informative?*				
System A		Neutral		System B
DialoGPT (345M)	3490 (**77%**)	206 (5%)	861 (19%)	PersonalityChat ****
DialoGPT (345M)	2474 (41%)	257 (4%)	3269(**54%**)	DialoGPT (345M, w/ MMI) ****
DialoGPT (345M)	3230 (**54%**)	362 (6%)	2408(40%)	DialoGPT (345M, Beam) *****
DialoGPT (345M)	2856 (**48%**)	303 (5%)	2841(47%)	DialoGPT (762M)
DialoGPT (345M)	2722 (45%)	234 (4%)	3044(**51%**)	Human response ****
DialoGPT (345M, w/ MMI)	3011 (**50%**)	234 (4%)	2755(46%)	Human response **

Human-like: *A and B, which is more likely to be generated by human rather than a chatbot?*				
System A		Neutral		System B
DialoGPT (345M)	3462 (**76**)%	196 (4%)	899 (20%)	PersonalityChat ****
DialoGPT (345M)	2478 (41)%	289 (5%)	3233 (**54**%)	DialoGPT (345M, w/ MMI) ****
DialoGPT (345M)	3233 (**54**)%	340 (6%)	2427 (40%)	DialoGPT (345M, Beam) ****
DialoGPT (345M)	2847 (**47**)%	321 (5%)	2832 (47%)	DialoGPT (762M)
DialoGPT (345M)	2716 (45)%	263 (4%)	3021 (**50**%)	Human response ***
DialoGPT (345M, w/ MMI)	2978 (**50**)%	241 (4%)	2781 (46%)	Human response *

Table 7: Results of **Human Evaluation** for relevance, informativeness and human-response possibility, showing preferences (%) for our model (DialoGPT) vis-a-vis its variants and real human responses. Distributions are skewed towards DialoGPT with MMI, even when compared with human outputs. Numbers in bold indicate the most preferred systems. Differences in mean preferences are statistically significant where indicated (* $p \leq 0.01$, ** $p \leq 0.001$, *** $p \leq 0.0001$, **** $p \leq 0.00001$).

a propensity to express agreement with propositions that are unethical, biased or offensive (or the reverse, disagreeing with otherwise ethical statements). These are known issues in current state-of-the-art end-to-end conversation models trained on large naturally-occurring datasets. A major motive for releasing DIALOGPT is to enable researchers to investigate these issues and develop mitigation strategies. In no case should inappropriate content generated as a result of using DIALOGPT be construed to reflect the views or values of either the authors or Microsoft Corporation.

7 Conclusion

We have released an open-domain pre-trained model, DIALOGPT, trained on massive real-world Reddit dataset. The package consists of a distributed training pipeline and several pre-trained models that can be fine-tuned to obtain a conversation model on a moderately-sized customized dataset in few hours. DIALOGPT is fully open-sourced and easy to deploy, allowing users to extend the pre-trained conversational system to bootstrap training using various datasets, and as a building block to novel applications and methodologies. In future, we will investigate how to detect and control toxic generation, and leverage reinforcement learning to further improve the relevance of the generated responses and prevent the model from generating egregious responses.

8 Acknowledgements

We thank Yu Wang, Vighnesh Leonardo Shiv and Chris Quirk for helpful discussions.

References

M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov, A. Litinsky, V. Logacheva, A. Lymar, V. Malykh, M. Petrov, V. Polulyakh, L. Pugachev, A. Sorokin, M. Vikhreva, and M. Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, pages 4171–4186.

E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, S. Prabhumoye, A. W. Black, A. Rudnicky, J. Williams, J. Pineau, M. Burtsev, and J. Weston. 2019. The second conversational intelligence challenge (ConvAI2).

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

J. Gao, M. Galley, and L. Li. 2019a. Neural approaches to conversational AI. *Foundations and Trends® in Information Retrieval*.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019b. Jointly optimizing diversity and relevance in neural response generation. *NAACL-HLT 2019*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019c. Structuring latent spaces for stylized response generation. *arXiv preprint arXiv:1909.05361*.

M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568*.

Z. Hu, H. Shi, Z. Yang, B. Tan, T. Zhao, J. He, W. Wang, L. Qin, D. Wang, et al. 2018. Texar: A modularized, versatile, and extensible toolkit for text generation. *arXiv preprint arXiv:1809.00794*.

HuggingFace. 2019. PyTorch transformer repository. https://github.com/huggingface/pytorch-transformers.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *ACL*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 EMNLP System Demonstration*.

Oluwatobi Olabiyi and Erik T Mueller. 2019. Multi-turn dialogue response generation with autoregressive transformer models. *arXiv preprint arXiv:1908.01841*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. *arXiv preprint arXiv:1906.02738*.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2018. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

R. Sennrich, B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of ACL (Volume 1: Long Papers)*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*.

Vighnesh Leonardo Shiv, Chris Quirk, Anshuman Suri, Xiang Gao, Khuram Shahid, Nithya Govindarajan, Yizhe Zhang, Jianfeng Gao, Michel Galley, Chris Brockett, et al. 2019. Microsoft icecaps: An open-source toolkit for conversation modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics: System Demonstrations*, pages 123–128.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017a. Attention is all you need. In *Advances in NIPS 30*. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.

A Additional Details of Human Evaluation

Significance testing for the difference in means was performed using 10K bootstrap iterations. P-values are computed at $\alpha = 0.05$. The results are provided in Table 8. The differences between 345M (2) and 762M (6) models are not significant. Notably also, the differences between 345M model (2) and human response (1) are not statistically significant. The template for human evaluation is provided in Figure 2.

	System 1			System 2			Pairwise		
	Mean	Std	95% CI	Mean	Std	95% CI	Std	95% CI	P-Value
2 vs 1 Human-like	0.4527	0.0065	(0.4400, 0.4653)	0.5035	0.0065	(0.4909, 0.5162)	0.0127	(-0.0758, -0.0259)	0.0001
2 vs 1 Informativeness	0.4537	0.0065	(0.4410, 0.4663)	0.5073	0.0064	(0.4948, 0.5199)	0.0127	(-0.0785, -0.0287)	0.0000
2 vs 1 Relevance	0.4452	0.0064	(0.4326, 0.4577)	0.4693	0.0064	(0.4568, 0.4819)	0.0124	(-0.0485, 0.0002)	0.0552
2 vs 3 Human-like	0.7597	0.0064	(0.7473, 0.7723)	0.1973	0.0059	(0.1858, 0.2089)	0.0117	(0.5392, 0.5852)	0.0000
2 vs 3 Informativeness	0.7659	0.0063	(0.7536, 0.7783)	0.1889	0.0058	(0.1777, 0.2003)	0.0115	(0.5540, 0.5993)	0.0000
2 vs 3 Relevance	0.7200	0.1935	(0.7070, 0.7333)	0.1935	0.0067	(0.7070, 0.7333)	0.0117	(0.5034, 0.5493)	0.0000
2 vs 4 Human-like	0.4130	0.0063	(0.4005, 0.4253)	0.5388	0.0064	(0.5263, 0.5514)	0.0124	(-0.1504, -0.1016)	0.0000
2 vs 4 Informativeness	0.4123	0.0063	(0.3999, 0.4246)	0.5448	0.0064	(0.5323, 0.5575)	0.0124	(-0.1570, -0.1082)	0.0000
2 vs 4 Relevance	0.3965	0.0063	(0.3841, 0.4088)	0.5157	0.0064	(0.5031, 0.5281)	0.0122	(-0.1431, -0.0955)	0.0000
2 vs 5 Human-like	0.5388	0.0064	(0.5263, 0.5513)	0.4045	0.0063	(0.3921, 0.4169)	0.0125	(0.1098, 0.1587)	0.0000
2 vs 5 Informativeness	0.5383	0.0064	(0.5258, 0.5508)	0.4013	0.0063	(0.3890, 0.4137)	0.0124	(0.1127, 0.1611)	0.0000
2 vs 5 Relevance	0.5032	0.0064	(0.4906, 0.5157)	0.4000	0.0063	(0.3876, 0.4124)	0.0122	(0.079, 0.127)	0.0000
2 vs 6 Human-like	0.4745	0.0065	(0.4618, 0.4872)	0.4720	0.0064	(0.4596, 0.4846)	0.0125	(-0.0220, 0.0272)	0.8476
2 vs 6 Informativeness	0.4760	0.0064	(0.4634, 0.4887)	0.4735	0.0064	(0.4610, 0.4861)	0.0126	(-0.0221, 0.0273)	0.8449
2 vs 6 Relevance	0.4543	0.0065	(0.4417, 0.4671)	0.4497	0.0064	(0.4372, 0.4622)	0.0123	(-0.0193, 0.0289)	0.7066
4 vs 1 Human-like	0.4963	0.0064	(0.4838, 0.5090)	0.4635	0.0065	(0.4508, 0.4762)	0.0127	(0.0081, 0.0578)	0.0094
4 vs 1 Informativeness	0.5018	0.0064	(0.4894, 0.5144)	0.4592	0.0127	(0.0180, 0.0676)	0.0127	(0.0180, 0.0676)	0.0009
4 vs 1 Relevance	0.4785	0.0064	(0.4660, 0.4911)	0.4345	0.0065	(0.4218, 0.4472)	0.0123	(0.0199, 0.0682)	0.0005

Table 8: Human evaluation significance test. Bold results represent differences that are **NOT** statistically significant. Notation: 1 - Human response; 2 - DIALOGPT 345M; 3 - PersonalityChat; 4 - DIALOGPT 345M w/ MMI; 5 - DIALOGPT 345M Beam search; 6 - DIALOGPT 762M

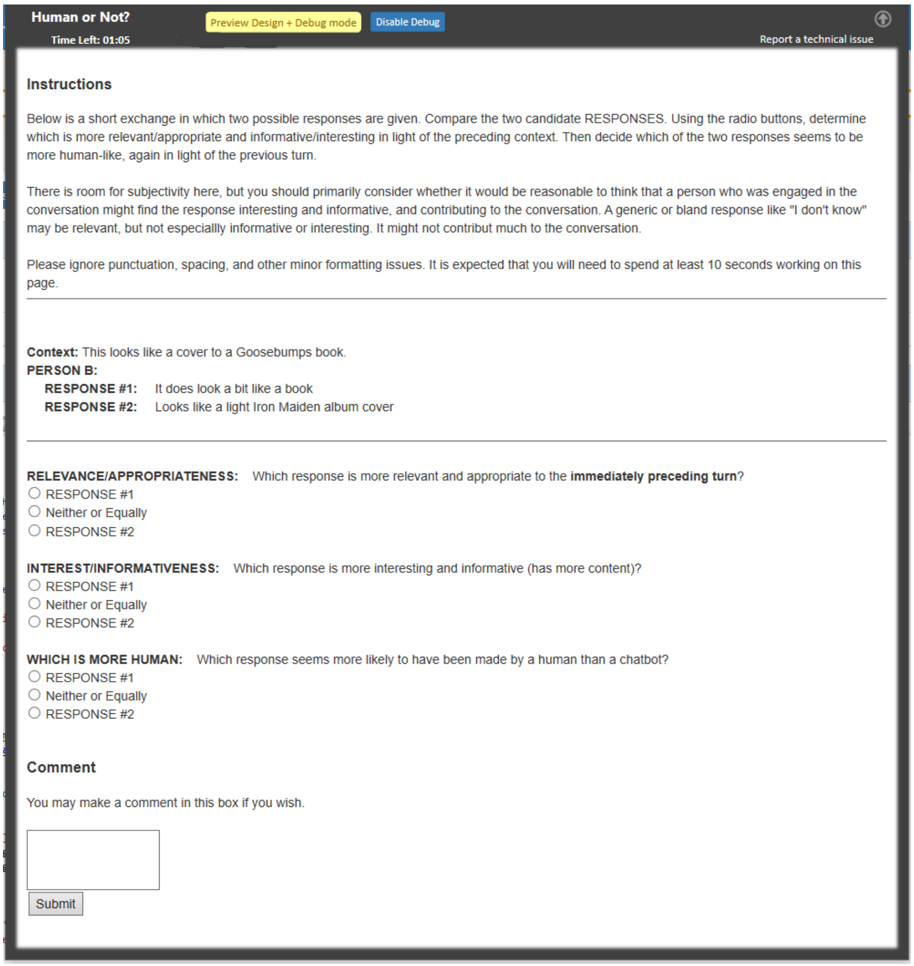

Figure 2: Human evaluation template