Personalized Chit-Chat

Junda An, Tiancheng Zheng, Ariel Xiao(yuweix)
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The main task of generative-based chatbot is to generate consistent, fluent and engaging response given the context, and this is a major focus in the field of natural language processing. In this report, we explore several state-of-the-art models for personalized chit-chat, build new models, and present our improved metric scores compared to the original papers.

GitHub Repository: https://github.com/JosephZheng1998/DialoGPT-Chat-Bot

1 Introduction

Generative-based Chatbots are becoming more and more prevalent nowadays in our daily lives. We need these Chatbots to perform daily duties such as answering questions for customer services and making smart phones more intelligent. These Chatbots are dialogue system that can be classified as two types: 1)open domain dialogue systems (Ritter et al., 2011) (Galley et al., 2015) (Serban et al., 2016) (Li et al., 2016) and 2)task-oriented dialogue systems. (Levin et al., 1997) (Wen et al., 2015) (Williams & Zweig, 2016) For the open-domain dialogue system often it does not set any limitation about the domains, while the task-oriented dialogue system focused on the present a consistent personality. In this way, it could gain the customers' trust and confidence. For Open-domain dialogue models can be divided into two classes: generation-based and retrieval-based. The generative models try to generate utterances conditioned on dialogue history and the retrieval-based model designed to select the best one that suit the situation from the prebuilt knowledge base. However, enforcing consistency, context-awareness and penalization is still a big challenge. In this report, we first explore several state-of-the-art models including P2 Bot, TransferTransfo, Lost In Conversation, then used DialoGPT, pretrained on Dialogue NLI and multi-task learning to further improve the performance for our Personalized Chit-Chat.

2 Background

2.1 Dataset Description

The dataset we used is the ConvAI2 dataset, which is based on the Persona-Chat dataset. The speaker pairs "each have assigned profiles coming from a set of 1155 possible personas (at training time), each consisting of at least 5 profile sentences, setting aside 100 never seen before personas for validation" (Dinan et al., 2019).

The detail of the dataset is presented in the table below:

	Examples	Dialogues	Personas
Training Set	131,438	17,878	1,155
Validation Set	7,801	1,000	100
Test Set	6,634	1,015	100

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2.2 Evaluation Metrics

There are three different automatic evaluation metrics: Perplexity, F1 and Hits@1/20.

1. Perplexity: a metric of text fluency which is computed as $\frac{1}{m}\sum_{i=1}^{m} logp(w_i)$ for sentence $\mathbf{w} = w_1, w_2, ..., w_m$. This metric is computed only for probabilistic generative models (Dinan et al., 2019).
2. F_1-score: $2 \cdot \frac{\texttt{precision}\cdot\texttt{recall}}{\texttt{precision}+\texttt{recall}}$. In the context of dialogue, precision is the fraction of words in the predicted response that are contained in the gold response, and recall is the fraction of words in the gold response that were in the predicted response. This can be computed for any model, retrieval-based or generative (Dinan et al., 2019).
3. Hits@1/20: hits@1/N is the accuracy of the next dialogue utterance when choosing between the gold response and $N-1$ distractor responses (here, we use $N = 19$). Distractor responses are random responses from the dataset. Any model that can assign a score to a given candidate utterance can compute this metric. Such a method could then in principle be used in a retrieval model to score retrieved candidates (Dinan et al., 2019).

3 Literature Review and Model

Open-domain dialogue models and generation-based models usually generate more diverse set of the answers so have better ability to adapt for different situation. However, it may have higher level of complexity of dependency and take longer time to train. For the other type of the model which is retrieval-based model would reply on the size and the accuracy of the dataset and knowledge base which would be pretty challenging and costly in the real world. For the later part of the report, we have the human-written dialogues providing neat examples of our daily communication and we would rebuild P^2 Bot, TransferTransfo, Lost In Conversation, then used DialoGPT, pretrained on Dialogue NLI and multi-task learning to further improve the performance for our Personalized Chit-Chat.

3.1 *P^2 Bot* (liu-etal-20 20-personachat)

P^2 Bot is proposed by Liu et al. in the paper "You Impress Me: Dialogue Generation via Mutual Persona Perception" in 2020. P^2 Bot is a "transmitter-receiver based framework with the aim of explicitly modeling understanding" (Liu et al., 2020). It "incorporates mutual persona perception to enhance the quality of personalized dialogue generation" (Liu et al., 2020) with the Transmitter and the Receiver. The Transmitter is responsible for dialogue generation while the Receiver is responsible for mutual persona perception. It achieves 82.5 (original:81.9) Hits@1 score, 18.77 (original:19.77) Perplexity, and 19.88 (original:15.12) F1 score.

3.2 *TransferTransfo* (Wolf et al., 2019)

TransferTransfo is a generative transformer model which is largely based on the Generative Pre-trained Transformer (Radford & Narasimhan, 2018). It has 12-layer with self-attention heads where every tokens only attends to the past tokens. The model is first pretrained on the BooksCorpus dataset (Zhu et al., 2015) and is finetuned on the PERSONA-CHAT dataset with augmented embeddings and multi-task learning. It achieves 82.1 Hits@1 score, 19.09 Perplexity, and 17.51 F1 score.

3.3 *Lost In Conversation(NeurIPS 2018)*

Lost In Conversation combines an encoder-decoder architecture based on a modified version of OpenAI GPT and a transfer learning approach to training, pretraining our model on a separate large dataset and then fine-tuning for the actual conversational datasets to solve the challenges of redundant, contrived, and extremely under-defined, and maybe all at the same for human conversations. One of the main challenges when we implement and improve this method is the version. Due to the maintenance issue, we spend a huge effort on building the docker image from scratch and setup with EC2-P4 instance, implement and modified ConvAI2 from ParlAI. We also uses multiple of the

epochs and adding additional training set and you could see the eventual result is better than the one in the paper. It achieves 18.1 Hits@1 score (original paper:17.1) and 18.6 (original paper:17.7) F1 score.

4 Novel Model

4.1 DialoGPT

Pretraining generates good results for both the TransferTransfo model and the Lost In Conversation model. It is worth to notice that the TransferTransfo model is pretrained on a corpus dataset and the Lost In Conversation is pretrained on DailyDialog which contains only 13,118 dialogues. We expect pretraining on a larger dialogue dataset would generate more consistent responses in the dialogue.

DialoGPT (Zhang et al., 2020) achieves a state-of-art performance in generating relevant, contentful, and context-consistent responses in the dialogue setting. Therefore, we use the DialoGPT that has already been trained on a very large dataset (147,116,725 dialogue instances, in total 1.8 billion words) as an encoder in our model. It would allow the model to capture the long-term dependency.

4.2 Pretrained on Dialogue NLI

One of the key goals in the personalized Chit-Chat task is to achieve consistency in the dialogue. For example, the persona statements of a bot include "I have two cats". If I ask him what are names of your pets and he answers "no, I do not have any pets", the response is inconsistent or contradictory. In order to allow the model to have a sense of natural language inference (NLI), a dataset called Dialogue NLI is proposed (Welleck et al., 2019). Each input of the dataset is generated by extracting two sentences from the Persona-Chat dataset to form a (utterance, persona) or a (persona, persona) pair. The label is the relation between these two sentence, including entailment, natural, and contradiction. Our model would be trained to classify each sentence pair to the correct class. The original paper (Welleck et al., 2019) shows that pretraining with the Dialogue NLI dataset, the KV-Mem baseline model achieves higher Hits@1 score. Therefore, we expect our model to achieve better performance after being pretrained with the NLI dataset.

4.3 Multi-task Learning

We adapt the multi-task learning scheme used in the TransferTransfo model. Therefore, the training of our model consists of two tasks. The first task is a language modeling during which we optimize the cross-entropy loss on the last softmax layer. The second task is to distinguish a correct next utterance and an incorrect next utterance, which is similar to the next sentence prediction in the BERT model. A set of incorrect utterances is randomly sampled and added to each training samples. A linear classifier is added to the last hidden layer to make the classification. The model is trained to maximize the log probability of the correct utterances. This additional classification task allows the model to capture more sense of a dialog rather than simply focusing on the next token.

The complete process of our method is 1) use DialoGPT as an encoder 2) pretrain our model with the Dialogue NLI dataset 3) use the multi-task learning scheme to finetune our model.

5 Experiments and Results

5.1 Finetuning Details

We experimented with the DialoGPT-small (486MB) and the DialoGPT-medium (823MB) models. We did not try the DialoGPT-large model (1.6GB) because it was too big to fit into the memory. We trained our models on an AWS p4d.24xlarge instance using the Deep Learning AMI with 8 A100-SXM4-40GB GPUs and CUDA version 11.0.

We trained the DialoGPT-small model for 3 epochs with the following hyperparameters.

Hyperparameter	Value
Number of candidates for training	4
Number of previous exchanges to keep in history	2
Batch size for training	2
Batch size for evaluation	2
Number of steps to accumulate gradients	4
Learning rate	6.25e-5
LM loss coefficient	2.0
Multiple-choice loss coefficient	1.0
Clipping gradient norm	1.0
Number of permutations of personality sentences	2

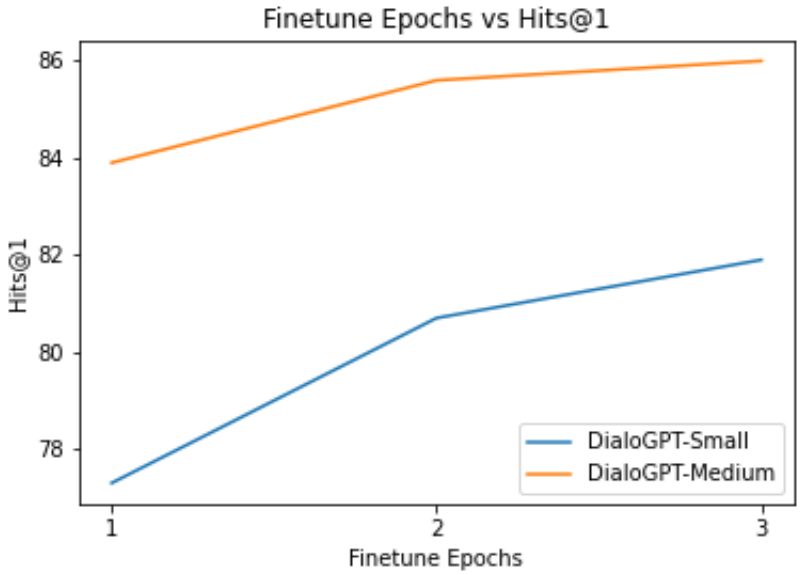
We trained the DialoGPT-medium model for 3 epochs with the following hyperparameters.

Hyperparameter	Value
Number of candidates for training	4
Number of previous exchanges to keep in history	2
Batch size for training	4
Batch size for evaluation	4
Number of steps to accumulate gradients	4
Learning rate	6.25e-5
LM loss coefficient	2.0
Multiple-choice loss coefficient	1.0
Clipping gradient norm	1.0
Number of permutations of personality sentences	2

We generated response by using beam size 4.

5.2 Results

	Hits@1	Perplexity	F1
DialoGPT-small	81.9	-	16.9
DialoGPT-medium	86.0	-	17.82
P^2 Bot	81.9	19.77	15.12
TransferTransfo	82.1	19.09	17.51
Lost In Conversation	17.1	-	17.77



From the Epochs vs Hits@1 plot, we can see that the metrics gradually converges as the fine-tuning epoch increases. In fact, in the fourth epoch, we start to notice overfitting in the model. Therefore, we only keeps the result of the first three epochs.

Our model, especially the DialoGPT-medium model, is clearly better than the other three state-of-the-art models mentioned before in terms of both Hits@1 and F1 metrics. Although it should be noted that our model is evaluated using the evaluating script provided by ConvAI2 dataset against its validation sets while the other three models are evaluated against the test set, which we do not have access to, our models still demonstrate a clear advantage over the other models.

We did not test on the perplexity because it took over ten hours to complete a single evaluation which would cost over $\$300$ dollars.

6 Error Analysis

In the real life dialog, humans rarely repeat themselves. However, when we experimented with the model, we observed that it tended to repeat themselves. For example, we observed "I like watching love movies" followed by "I love watching romantic movies". To measure the level of repetitions, we manually had 50 conversations with our model. If two responses conveyed the exact same meaning when it was not asked to repeat themselves, we denoted them as non-unique. As a result, 92% responses are unique. We expect the uniqueness rate to be close to 100% or 90% when humans have conversation. However, since the persona sentences are limited for each conversation, the model will repeat themselves more frequently if we want the model to be more personal (in other words, generate responses that describe its persona).

In addition, our model tended to ask questions. When we experimented with this model, it either generated responses starting with "what, how, when, where, how" or ending the sentence with a question very often. In order to measure how frequently our model tended to ask questions, we manually had 50 conversations with our model. The model asked 38 questions. (Note that each conversation may consist of multiple sentences or responses). Too many questions would make the conversation unnatural.

7 Future Work

For the repetition error, we will mitigate it by developing a metric that measures the tendency of the model to repeat themselves and optimize it during training.

For the too-many question issue, we will penalize the question marks if it asks questions too frequently in the training process. In this way, will we balance the question-answering and the question-asking in the conversation.

As DialoGPT-medium outperforms DialoGPT-small, we expect DialoGPT-large will generate better results. If we are given more budget and more computation power, we will try to use DialoGPT-large as an encoder to improve our model.

References

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., Williams, J., Pineau, J., Burtsev, M., & Weston, J. (2019). The second conversational intelligence challenge (convai2).

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., & Dolan, B. (2015). DeltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 445–450. https://doi.org/10.3115/v1/P15-2073

Levin, E., Pieraccini, R., & Eckert, W. (1997). Learning dialogue strategies within the markov decision process framework. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 72–79. https://doi.org/10.1109/ASRU.1997.658989

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model.

Liu, Q., Chen, Y., Chen, B., Lou, J.-G., Chen, Z., Zhou, B., & Zhang, D. (2020). You impress me: Dialogue generation via mutual persona perception. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Ritter, A., Cherry, C., & Dolan, W. B. (2011). Data-driven response generation in social media. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 583–593. https://www.aclweb.org/anthology/D11-1054

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models.

Welleck, S., Weston, J., Szlam, A., & Cho, K. (2019). Dialogue natural language inference.

Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D., & Young, S. (2015). Semantically conditioned lstm-based natural language generation for spoken dialogue systems.

Williams, J. D., & Zweig, G. (2016). End-to-end lstm-based dialog control optimized with supervised and reinforcement learning.

Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, *abs/1901.08149*. http://arxiv.org/abs/1901.08149

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020). Dialogpt: Large-scale generative pre-training for conversational response generation.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *The IEEE International Conference on Computer Vision (ICCV)*.