Is Your Chatbot Perplexing?: Confident Personalized Conversational Agent for Consistent Chit-Chat Dialogue

Young Yun Na[1,*][a], Junekyu Park[2,*][b] and Kyung-Ah Sohn[2,3,†][c]
[1] *Department of Media, Ajou University, Suwon, Gyeonggi, Republic of Korea*
[2] *Department of Artificial Intelligence, Ajou University, Suwon, Gyeonggi, Republic of Korea*
[3] *Department of Software and Computer Engineering, Ajou University, Suwon, Gyeonggi, Republic of Korea*

[*] *These authors contributed equally,* [†] *Corresponding author*

Keywords: Chatbot, Dialogue, Personalized, Confidence, Natural Language Processing.

Abstract: Chatbots are being researched and employed not only in academic settings but also in many fields as an application. Ultimately, conversational agents attempt to produce human-like responses along with dialogues. To achieve this goal, we built a novel framework that processes complex data consisting of personalities and utterances and fine-tuned a large-scale self-attention-based language model. We propose a consistent personalized conversational agent(*CPC-Agent*) for the framework. Our model was designed to utilize the complex knowledge of a dataset to achieve accuracy and consistency. Together with a distractor mechanism, we could generate confident responses. We compared our model to state-of-the-art models using automated metrics. Our model scored 3.29 in perplexity, 17.59 in F1 score, and 79.5 in Hits@1. In particular, the perplexity result was almost four times smaller than that of the current state-of-the-art model that scored 16.42. In addition, we conducted a human evaluation of each model to determine its response quality because the automatic evaluation metrics in dialogue tasks are still considered insufficient. Our model achieved the best rates from the voters, which indicated that our model is adequate for practical use.

1 INTRODUCTION

The need for automatic-response generation in machine-to-human interaction is increasingly being emphasized. However, few studies have achieved a sensible quality of auto-generated responses. Conventional chatbots, which are human interactive response-generating machines, are built based on tree structures that can only produce pre-scripted responses. To train chatbots that generate abstractive and human-like responses, recent studies have employed not only deep-learning approaches but also different dataset formats. Since the appearance of the transformer-based models (Vaswani et al., 2017), such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), deep-learning-based language models have been successful in most natural language processing(NLP) fields. However, the dialogue-generation task is considered one of the tasks that have yet to be mastered. Recent deep-learning-based chatbot models are trained with an utterance text dataset. However, training the model using only uttered text can make it generate style-impersistent responses and suffer from the lack of long-term context information.

In this paper, we propose a contextually consistent GPT-based chatbot model that uses PERSONA-CHAT dataset (Zhang et al., 2018) that contains an additional personality as consistency information. An example of a dialogue is shown in Figure 1. For experiments, we evaluated the model using F1, perplexity(PPL), and Hits@1 as automated metrics and conducted a human evaluation on the quality of the auto-generated responses. The main contributions of our model are as follows:

- We built a GPT-based deep trainable conversational agent that could generate abstractive responses and engage in interactive dialogue.
- We fine-tuned a dialogue-dedicated model that was pre-trained using a large-scale conversational dataset.
- We carefully designed a training mechanism to process personality and utterance input and com-

[a] https://orcid.org/0000-0001-7051-3766
[b] https://orcid.org/0000-0003-2984-9027
[c] https://orcid.org/0000-0001-8941-1188

Na, Y., Park, J. and Sohn, K.
Is Your Chatbot Perplexing?: Confident Personalized Conversational Agent for Consistent Chit-Chat Dialogue.
DOI: 10.5220/0010385312261232
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 2*, pages 1226-1232
ISBN: 978-989-758-484-8

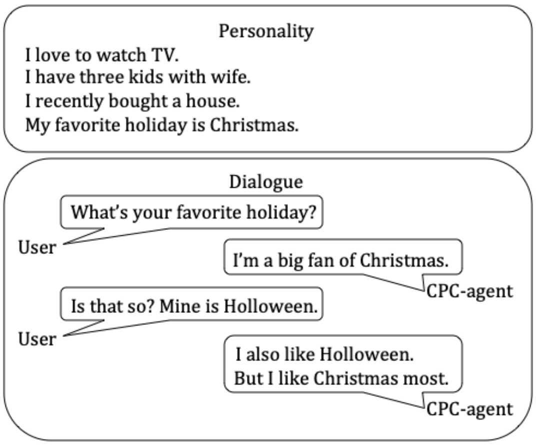

Personality
I love to watch TV.
I have three kids with wife.
I recently bought a house.
My favorite holiday is Christmas.
Dialogue
What's your favorite holiday?
User
I'm a big fan of Christmas.
CPC-agent
Is that so? Mine is Holloween.
User
I also like Holloween.
But I like Christmas most.
CPC-agent

Figure 1: *CPC-Agent* chit-chat sample. *CPC-Agent* is a chatbot that generates responses when a user inputs sentences. it can also keep track of the conversation history. As a result, *CPC-Agent* can generate consistent and personalized sentences.

pared our model to the state-of-the-art models through automatic evaluation and human evaluations. Consequently, our model outperformed the state-of-the-art models in terms of PPL score and human evaluation.

The rest of this paper is organized as follows. In section 2, background of natural language generation and works related to the dialogue task are reviewed. Section 3, describes the details of the model framework proposed. Section 4, discuss the personalized dataset and model analysis using both automatic metric and human evaluation. Lastly, the conclusion comes in section 5.

2 BACKGROUND

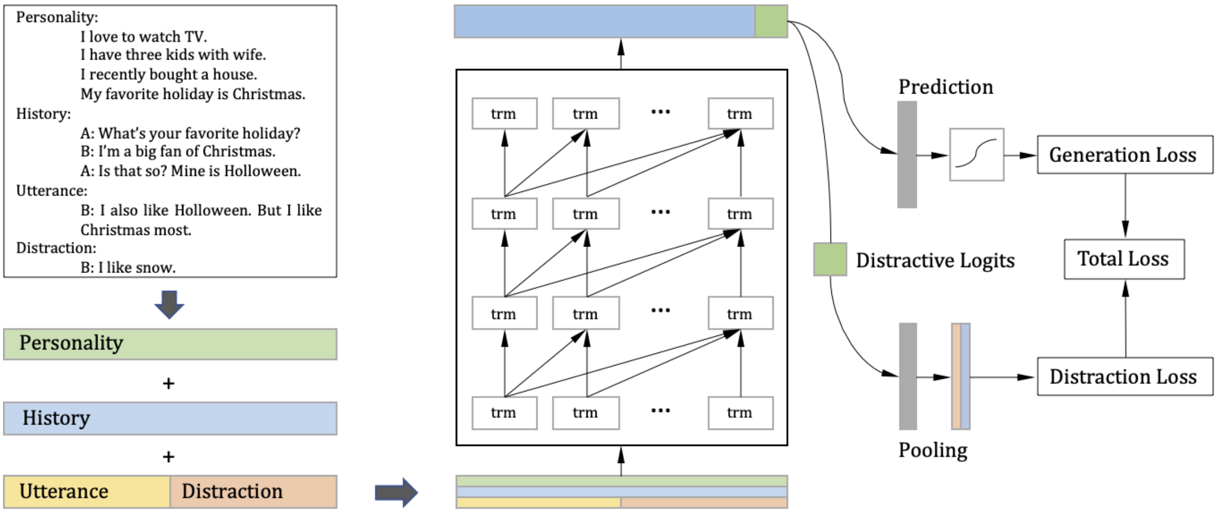
Deep learning-based language models in a dialogue task, are trained to generate interactive responses. Language models are mainly divided into models before and after the appearance of the transformer. The models before the transformer are mainly based on recurrent neural networks (RNNs), such as LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014). By employing the large capacity of RNN-based models, (Vinyals and Le, 2015) demonstrated acceptable responses in open-domain settings and (Serban et al., 2016; Miao et al., 2016; Sordoni et al., 2015) utilized the latent state from RNN models to generate abstractive responses.

More recent language models that adopted the transformer architectures have demonstrated dominant performance in most NLP tasks. (Mazaré et al., 2018) incorporated additional data from the 'Reddit' dialogue chit-chat to train a transformer-based model. (Wolf et al., 2019a) proposed a fine-tuned transformer-based language model and illustrated extended generative experiments. In the case of an open-domain dialogue, these models are incapable of generating accurate responses because there is no fixed topic across the conversation.

To further train the conversational model, (Zhang et al., 2018) presented a problem in which the dataset is the one that needs to be handled. A novel PERSONA-CHAT dataset was also proposed. In contrast to a mere utterance dataset, the PERSONA-CHAT dataset contained additional personality of each speaker, which could make the model generate consistent responses. To facilitate the personality information, (Golovanov et al., 2020) fine-tuned the GPT model pre-trained on manually labeled 'DailyDialog' using transfer learning. P^2 Bot, which was proposed by (Liu et al., 2020), underwent an additional reinforcement learning approach. The P^2 Bot processed the personality information using the BERT architecture that learned from word-piece representation and processed the utterance-based conversational information using the GPT architecture which learned from sentence-piece representation. This separate processing pipeline required the model to share the context through reinforcement learning. The encoding methods in each processing pipeline also differed. Thus, sharing the context of each propagated input was difficult. In contrast to P^2 Bot, $TransferTransfo$ model which was proposed by (Wolf et al., 2019b), employed a different approach to process the personalities along with the utterances. $TransferTransfo$ obtained a higher accuracy by training the GPT model with a distractor by using other candidate utterances but simply used the embedding form of persona with utterances of agents as an input. However, this form of input did not consider the dialogue history. Thus, it still suffers from limitation in terms of remembering long-term memory.

3 MODEL STRUCTURE

We propose a framework to train the generative model for a dialogue task. Our model is designed to process complex data information in a sophisticated and efficient manner. To profoundly train the model, we adopt the GPT-based model which utilizes a self-attention mechanism. Overall, we train the model through two combined losses using the property of distinct contexts gathered in the encoded embedding.



Personality:
I love to watch TV.
I have three kids with wife.
I recently bought a house.
My favorite holiday is Christmas.
History:
A: What's your favorite holiday?
B: I'm a big fan of Christmas.
A: Is that so? Mine is Holloween.
Utterance:
B: I also like Holloween. But I like Christmas most.
Distraction:
B: I like snow.

Figure 2: Overall structure of *CPC-Agent*. The *CPC-Agent* is trained using the PERSONA-CHAT dataset, which contains additional personality information. The inputs are the concatenated form of personalities, history of utterances, and utterances on both true and false responses that should be generated. The input is processed using the GPT-based model with two combined losses; generation and distraction losses. The training proceeds as a total loss plus generation loss and distraction loss. training The blocks in the model are transformer units.

This structure is shown in Figure 2. The *CPC-Agent*[1] is implemented using three different components as explained below.

3.1 Training with Personality

We propose to train the model using a complex dataset consisting of personalities and utterances. Personalities can be defined as the persona of each agent and the utterances represent the dialogue of two different agents A and B. In the case of $TransferTransfo$, used input that can be formulated as a concatenation of each vector which does not consider about history; $[p, u_A, u_B]$.

Unlike $TransferTransfo$, the $CPC-Agent$ model takes the utterance history into account as additional input information and processes the whole input using an integrated single model to facilitate context sharing. To fully understand and process the entire context from the given information, because the input format is x_n with n^{th} response generation, we consider the persona and history of dialogue and utterance of the agent B as one response using the generated sentence and distraction d_n which we will explain in Section 3.3. This input format is described as follows:

$$x_n = [p, h_n, u_{B_n}, d_n],$$
$$h_n = \sum_{i=1}^{n-1} [u_{A_i}, u_{B_i}], \tag{1}$$

where n is the number of pairs of utterances from the conversational agents so far, and h denotes the history of conversation before n. History provides the whole text of the dialogue sequences. By processing the given input format, the model can take the best advantages of the persona and all of the conversational information.

3.2 Generative Loss

The GPT-based models have been showing the best performance in most NLP fields. In particular, they demonstrated remarkable results in the field of text generation. As a result, DialoGPT was introduced by (Zhang et al., 2019) as a pre-trained model to generate interactive responses. For the structure of the model, we fine-tuned the model of the DialoGPT structure with the pre-trained weights. By utilizing the advantages of the models that used the self-attention mechanism, we were able to introduce x_n into a combined form of embedding from p, h_n and u_{B_n} by using special tokens $< SP1 >$ and $< SP2 >$ as separators. For training, the *CPC-Agent* processed the input x_n using a self-attention mechanism with the embedding dimension size dim. The model obtained output $o_{n_i} \in R^{dim}$, which was a vector for a word i with the context from the persona, dialogue history, and utterance of agent B. The next layer was a regressing layer for generation that yielded $q_i \in R^v$, where v is the size of the vocabulary. Finally, the sigmoid values are taken from a regression distribution q_i, and the k highest values are extracted from the probability

[1] github link : https://github.com/fightnyy/CPC-Agent

distribution. Denoting M as the number of generated words to consist $dialogue_n$ and N as the number of generated $dialogue_n$ to consist $Dialogue_n$, the objective function for generating the complete dialogue is as follows:

$$\begin{aligned} P(Dialogue) &= \prod_{n=1}^{N} P(dialogue_n|x_n), \\ P(dialogue_n|x_n) &= \prod_{i=1}^{M} P(pred_i|x_{n_i}), \end{aligned} \tag{2}$$

The reason for extracting the final output in the probability distribution of top k rather than extracting the highest one is that the model was designed to look and learn from the top k different candidates. Finally, we used the cross-entropy loss to train the model in which generative loss, L_{Gen} is defined as:

$$L_{Gen} = \frac{-1}{M} \sum_{i=1}^{M} [\log pred_{g_i} \cdot label_{g_i}], \tag{3}$$

where $pred_{g_i}$ is the next prediction of the model and $label_{g_i}$ as is the target word to generate at iteration i.

3.3 Distraction Loss

In the open-domain-dialogue tasks, a problem came up when sentences that have nothing to do with the previous conversations were generated. Inspired by (Zhang et al., 2018), we added distracting factors during the training to resolve the inconsistency issue by training *CPC-Agent* to distinguish the distraction and true utterance. For the distraction calculation, we used the binary cross-entropy loss as follows:

$$\begin{aligned} L_{Distract} = \frac{-1}{M} \cdot \sum_{i=1}^{M} [label_{d_i} \cdot \log pred_{d_i} \\ + (1 - label_{d_i}) \cdot \log(1 - pred_{d_i})], \end{aligned} \tag{4}$$

where $pred_{d_i}$ is the model prediction of whether u_B and d_n are either distraction or real response utterance, and $label_{d_i}$ is the ground truth of the prediction. For the total loss L_{total} for training the whole learning sequence, the model is trained by the combined generation and distraction losses as follows:

$$L_{total} = L_{Gen} + \alpha \cdot L_{Distract}, \tag{5}$$

where α is a hyperparameter as a regulating factor. From these generation and distraction losses, we could train the model using complex input information.

4 EXPERIMENT

Chatbot models for chit-chat go through the phenomenon of getting lost and generating irrelevant responses during a spoken conversation process without a particular topic. For consistency of conversation, the PERSONA-CHAT dataset was proposed by (Zhang et al., 2018), which could personalize the model when generating responses. Therefore, we trained our model using a PERSONA-CHAT dataset. The PERSONA-CHAT dataset was first used for training the model in ConvAI2 Competition, which created the PERSONA-CHAT dataset benchmark data, and we were able to compare our model with the scores claimed by the models trained using the persona data. The PERSONA-CHAT dataset consisted of 1,155 personalities where each described at least five sentences. For training, a total of 162,024 utterances in 10,907 dialogues were used. For testing, a total of 15,024 utterances over 968 dialogues were used.

To evaluate the effectiveness of our approach, we conducted automatic and human evaluations. In the automatic evaluation, we tested the models with multiple automatic metrics. In addition, as in the human evaluation method of (Liu et al., 2020), we conducted an experiment with 200 voters above-university educational level to determine how consistently *CPC-Agent* would respond to the persona. We also examined how the automatic evaluation results varied by adjusting the values of hyperparameter α. Since test-set of PERSONA-CHAT dataset is unavailable all test is conducted by using dev-set of PERSONA-CHAT.

4.1 Automatic Evaluations

We used three different automatic metrics; F1 score, perplexity(PPL), and Hits@1 for automatic evaluations of the models. The perplexity score represents the measurement of the negative log likelihood of the correct sequence output by the model. This score indicates the number of words the model considers for the prediction of the next words. The Hits@1 score represents the probability-based measure that ranks the real response from the highest according to the model. The F1 score represents the harmonic mean of word-level precision and recall.

The comparison results of *CPC-Agent* with those of the prior models tested on the same data under the same setting are listed in Table 1. Our model approach was comparable to all the baselines and advanced models. Particularly we achieved a new state-of-the-art performance on PPL scoring at 3.29, achieving a huge improvement, which was the low-

Table 1: Automatic evaluation of the models tested using the PERSONA-CHAT dataset; perplexity is denoted as PPL. Seq2Seq + Attention denotes the baseline in ConvAI2 Competition. All models are evaluated in terms of perplexity, F1, and Hits@1 as generated-response-quality measurements. Our model outperforms every other model in terms of perplexity.

Model	PPL(Perplexity)	F1(%)	Hits@1(%)
p^2 BOT	15.12	**19.77**	81.9
TransferTransfo	17.51	19.09	**82.1**
Lost In Conversation	17.3	17.79	17.3
Seq2Seq + Attention	35.07	16.82	12.5
CPC-Agent(**ours**)	**3.29**	18.89	79.5

Table 2: Human evaluation of the models tested using PERSONA-CHAT dataset. Each output of the models is rated in four different degrees, namely, basic, good, better, and excellent, through human voting. The average scores are calculated as averaged ratings from each degree. Our model scored the best in terms of excellent grade and average of ratings.

Model	Voting rate (%)				Score
	Basic: 1	Good: 2	Better: 3	Excellent: 4	Average
p^2 BOT	18.9	26.3	28.6	26.2	2.621
TransferTransfo	**41.7**	25.3	**28.7**	4.3	1.956
Lost In Conversation	26.3	**48.7**	22.0	3.0	2.017
CPC-Agent(**ours**)	19.1	26.5	23.5	**30.9**	**2.662**

est among the compared models. It was more than 10 times smaller than the baseline PPL score. Apart from the other metrics, the reason that PPL appeared particularly good was first, a large-scale pre-trained representative mode was used, and second, the model itself was more robust with the insertion of a distractor during the training. This result verified that our model exhibited better confidence in deciding the next words. The F1 score of our model was 18.89, which was competitive with that of the comparing models, but performed better than the *Lost In Conversation* and baseline models. In terms of Hits@1, our model scored 79.5, which largely differed from that in *Lost In Conversation* and the baseline models. It did not outperform the score of 81.9 from p^2 Bot and 82.1 from *TransferTransfo* but the differences were an acceptable limit.

4.2 Human Evaluation

In particular, in the dialogue task, even if the model obtained a good score on the automatic metric, it does not mean the performance of the model was much better. Automatically evaluating polyphonic or fluency was very difficult. For example, “I love my dog” and “puppy is liked by me” mean almost the same thing for a human. However, neither F1 nor Hits@1 considered them to have the same meaning. As an example of this, Table 1 lists that *TransferTransfo* and *Lost In Conversation* occupied the first and second places, respectively, in the NeurIPS 2018 dialog competition ConvAI2. *TransferTransfo* was rated as a better model than *Lost In Conversation* in all scores. but the human evaluation of these models showed opposite results. Hence, we conducted an additional human evaluation on our model.

Table2 lists the human evaluation of the compared models using the PERSONA-CHAT dataset. We requested people who had a higher university education level to evaluate our model. In the survey, each option was labeled as follows:

- Basic(1): The response is good only in terms of grammar and sentence structure.
- Good(2): Not only B’s response is consistent but also the response is coherent with the context.
- Better(3): B’s response is coherent meanwhile interesting and informative, instead of just a simple response like ”Yes.”
- Excellent(4): The response of B is consistent with the persona.
- Average: Average of each score multiplied by the rating portion.

TransferTransfo received the most votes of 41.7% on the “basic(1)” which indicated the generated responses from the model might be good but inconsistent. In terms of “good(2)”, the score of *Lost In Conversation* was 48.7% of votes, which was highest among the compared models. p^2 Bot did not obtain the highest in any of the options. In terms of the “excellent(4)” option, our model obtained the 30.9% of votes from the survey, which implied that the responses generated by our model were consistent with the personality of the chatbot. Therefore, the best model for the PERSONA-CHAT dataset criterion was closer to “excellent(4)”. Thus, our model demonstrated performance beyond p^2 Bot with a significant difference. The “average” score also illustrated that the performance of our model exceeded that of the other compared models, which meant that human felt most natural when they interacted with *CPC-Agent*.

4.3 Distraction Factor Adjustment

We recorded the score of the automatic metric according to the change in the α value to evaluate the change in the model performance with respect to the weight of the model influence of $L_{Distract}$. The value of α was set to 0.0(never used), 0.5, 1.0, 3.0, 5.0.

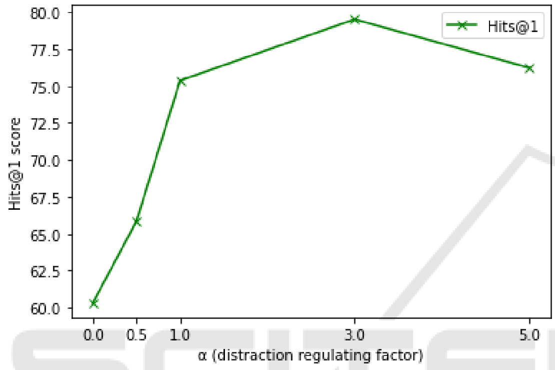


Figure 3: Hits@1 score curve along different each α value of our model. The x-axis represents the alpha value. The y-axis shows the score for Hits@1. The score represents the maximum when the α value is 3.0.

Figure 3 shows that the larger the α value is, the better is the result of Hits@1 until it reaches the maximum. When the α values were exactly 0.0, 0.5, 1.0, 3.0, and 5.0, the Hits@1 values were 60.32, 65.85, 75.37, 79.51, and 76.23, respectively. In particular, even if our model did not use a distractor, we could observe that the performance was good with a large difference between the existing attention mechanism and *Lost In Conversation*. The best performance was achieved when the α value was set to 3.0. However, it decreased when α value exceeded 3.0.

Figure 4 shows that PPL was most affected by the α value. In the case of the F1 score, the difference was insignificant. When the alpha values were 0.0, 0.5, 1.0, 3.0, and 5.0, the values of PPL were 20.41, 18.23, 5.65, 3.29 and 4.23 respectively, and the F1 values were 16.75, 16.90, 17.06, 18.89 and 17.23 respectively. Even when the distractor was not used, it easily surpassed the existing baseline. In particular, when α was 1.0, PPL exhibited the best performance improvement, and when α was 3.0, it demonstrated a maximum value. However, when it exceeded 3.0, the performance decreases. In terms of F1, it appeared that the influence of distractor is smaller than the other two metrics. However, the best performance was the same when α was 3.0.

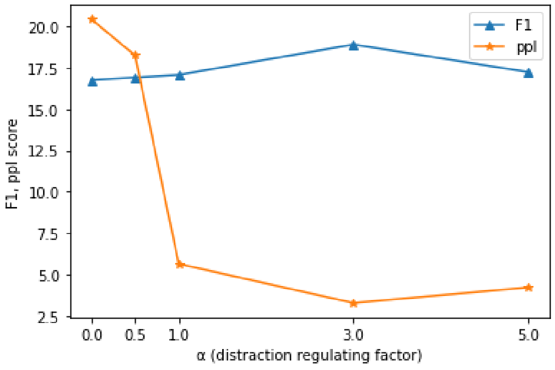


Figure 4: F1 and perplexity score curve along each different α value of our model. The x-axis represents the α value. The y-axis shows the perplexity and F1 scores. Both scores represent the maximum when the α value is 3.0.

5 CONCLUSIONS

We took advantage of the pre-trained representative model, which is known to perform well in the NLP tasks, and proposed a new GPT-based chatbot model that could serve as a consistent personalized conversational agent. By using a distractor mechanism in the model, for automatic evaluation, PPL demonstrated a large performance difference from the existing state-of-the-art models. Our model also demonstrated the best performance in human evaluation. The F1 and Hits@1 scores of our model showed performance comparable to the state-of-the-art models. Overall, our model can be considered as a state-of-the-art model on a conversational agent trained using personality information. However, we have not achieved weight reduction of the model. During the experiment, we found out that the number of parameters of transformer-based model (e.g, GPT, BERT) was too large, and in a general environment, it would be difficult to upload only the GPT model itself, even if not training GPT. In our future studies, in order to alleviate the model size problem, a lightweight model such as distill-gpt2 or distilled BERT or LSTM will be used to create a model whose parameters are much lighter than those of the current model.

ACKNOWLEDGEMENTS

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC

(Information Technology Research Center) support program(IITP-2020-2018-0-01431) and also under the National Program for Excellence in SW (2015-0-00908) both supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Golovanov, S., Tselousov, A., Kurbanov, R., and Nikolenko, S. I. (2020). Lost in conversation: A conversational agent based on the transformer and transfer learning. In *The NeurIPS'18 Competition*, pages 295–315. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Liu, Q., Chen, Y., Chen, B., Lou, J.-G., Chen, Z., Zhou, B., and Zhang, D. (2020). You impress me: Dialogue generation via mutual persona perception. *arXiv preprint arXiv:2004.05388*.

Mazaré, P.-E., Humeau, S., Raison, M., and Bordes, A. (2018). Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Miao, Y., Yu, L., and Blunsom, P. (2016). Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2016). A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., and Nie, J.-Y. (2015). A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019a). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2019b). Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.