Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge

Yoonna Jang[1]*, Jungwoo Lim[1]*, Yuna Hur[1]*, Dongsuk Oh[1], Suhyune Son[1], Yeonsoo Lee[2], Donghoon Shin[2], Seungryong Kim[1†], and Heuiseok Lim[1†]

[1]Department of Computer Science and Engineering, Korea University
[2]Language AI Lab, NCSOFT
{morelychee, wjddn803, yj72722, inow3555, ssh5131, seungryong_kim, limhseok}@korea.ac.kr
{yeonsoo, dhshin}@ncsoft.com

Abstract

Humans usually have conversations by making use of prior knowledge about a topic and background information of the people whom they are talking to. However, existing conversational agents and datasets do not consider such comprehensive information, and thus they have a limitation in generating the utterances where the knowledge and persona are fused properly. To address this issue, we introduce a *call For Customized conversation* (FoCus) dataset where the customized answers are built with the user's persona and Wikipedia knowledge. To evaluate the abilities to make informative and customized utterances of pre-trained language models, we utilize BART and GPT-2 as well as transformer-based models. We assess their generation abilities with automatic scores and conduct human evaluations for qualitative results. We examine whether the model reflects adequate persona and knowledge with our proposed two sub-tasks, persona grounding (PG) and knowledge grounding (KG). Moreover, we show that the utterances of our data are constructed with the proper knowledge and persona through grounding quality assessment.

Introduction

A person who is asked by a vegetarian to suggest a restaurant in New York City would not usually recommend Wolfgang's Steakhouse. When people give information to others, they consider the background of the person whom they are talking to. Following this manner of humans' conversation, a conversational agent's ability to have a conversation with *customized answers* from prior knowledge and user's personal information is crucial for satisfying the users. For example, as exemplified in Figure 1, the answer that considers both the user's persona and knowledge is much more attractive as well as informative.

Research for human-machine dialog has achieved significant success recently, owing to the advance of diverse dialog datasets (Adiwardana et al. 2020; Zhang et al. 2019b; Shuster et al. 2019; Li et al. 2017; Lowe et al. 2015) and pre-trained language models (Raffel et al. 2019; Clark et al. 2020; Brown et al. 2020). Despite the remarkable success, the model's ability to give knowledge-grounded answers reflecting user's personal information remains largely limited.

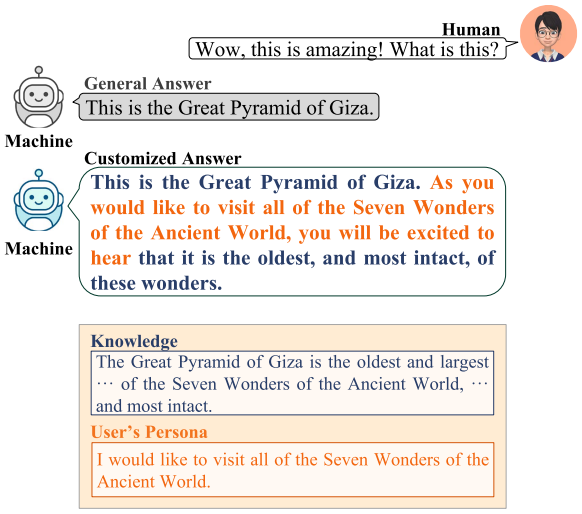

Figure 1: Objective of FoCus dataset. In contrast to the general answer, which only gives basic information, the machine's answer of FoCus dataset is more knowledgeable and customized, reflecting both knowledge and persona.

There exist several datasets and models that consider the user's persona, such as preference, interest or experience (Majumder et al. 2020; Xu et al. 2020; Wu et al. 2019; Zhang et al. 2018; Rashkin et al. 2018; Shuster et al. 2018; Li et al. 2017; Joshi, Mi, and Faltings 2017), which contributes to building an agent that can talk about the user's feelings and interests. Though the dialog agent can access to the persona, the absence of knowledge often limits its ability of generating answers with specialized knowledge.

Meanwhile, to build a dialog agent that generates more knowledgeable answers, datasets with the informative answers have been released (Dinan et al. 2018; Zhou, Prabhumoye, and Black 2018). In these datasets, the dialog agents learn to retrieve the required knowledge from the document. However, these datasets do not consider the user's persona,

*These authors contributed equally.
†These authors are corresponding authors.

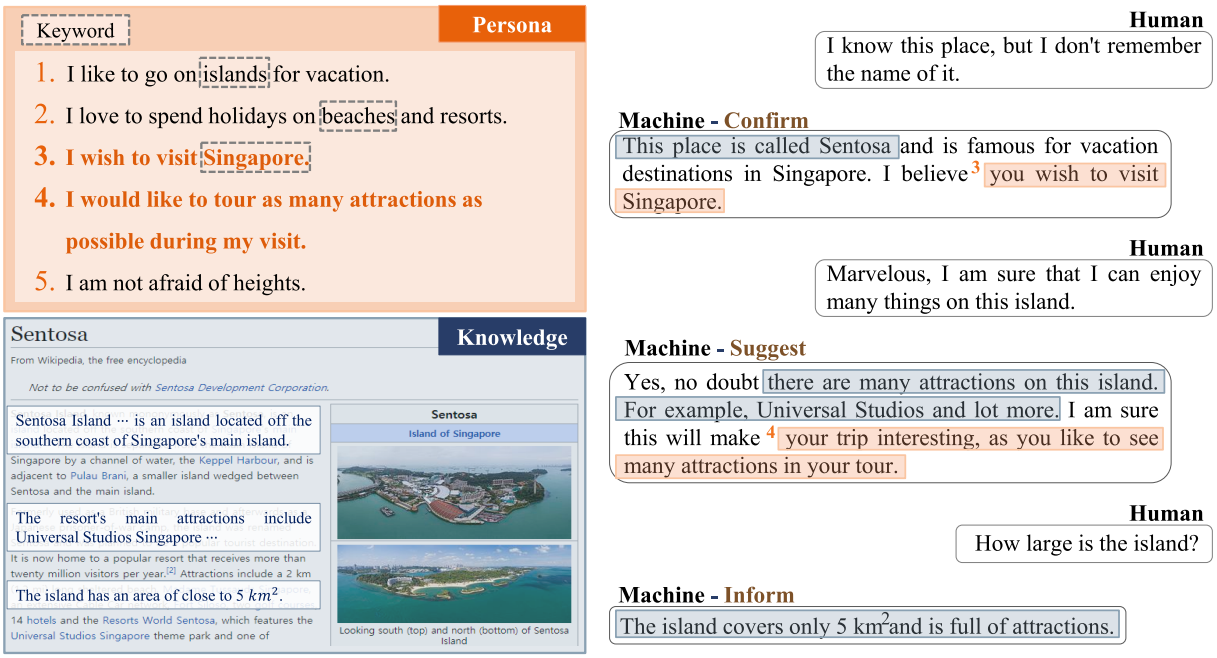

Figure 2: Example dialog between Human and Machine in FoCus dataset. The Human first asks about the landmark and the Machine then generates the answer considering the user's persona and Wikipedia knowledge. Answers can be made only with Wikipedia knowledge or both persona and Wikipedia knowledge. For instance, the third answer provides information about the size of the island only with knowledge. However, the second answer reflects both persona and knowledge.

which restrict generating customized answers. Providing a large amount of knowledge without considering the user's background may result in giving the user useless information because people may need different types of knowledge, depending on their interests.

For the ability to make use of both persona and knowledge, there have been a few attempts to blend them (Smith et al. 2020; Roller et al. 2020). However, they merely stitch up the existing datasets, thus the models process only one source at a time, not both of them. Little work had been done on fusing the persona and knowledge into the utterances, thus there could not be sufficient conditions to build customized and intelligent conversational agents.

In this work, we introduce a new dataset, *call For Customized conversation* dataset[1] (called FoCus), that supports knowledge-grounded answers that reflect user's persona. One of the situations in which people need different types of knowledge, based on their preferences, occurs when they travel around the world. As the knowledge of the landmark encompasses the range of history, design, structure, usage, tourism, and geological information, the diversity of the knowledge ensures. Inspired by this situation, we built a dataset where the agent informs the knowledge about the geographical landmark considering the user's persona.

Our contributions are as follows:

- We present the FoCus dataset in which the utterances contain both knowledgeable and customized answers for the first time.
- We propose the baseline generative models trained on our dataset and evaluate them with the automatic scores and conduct human evaluation in respect to the generation abilities.
- We provide two sub-tasks to measure the grounding ability, such as persona grounding (PG) and knowledge grounding (KG).

FoCus Dataset

To cover the diverse domain of a specific topic, we put the dialog under the setting of talking about Wikipedia knowledge on geographical landmarks. As the document of given landmarks provides various information of diverse domain, our dataset is well applicable to situations where the specialized knowledge is required. In this section, we describe the data collection process and analysis of the collected data. Also, we show three types of customized answers observed in our dataset.

Dataset Creation

We collected the conversations about the geographical landmark guidance through Amazon Mechanical Turk (AMT)[2]. For the topic of dialogs, we selected a landmark from Google Landmarks Dataset v2 (GLDv2) (Weyand et al. 2020). There are 5,316 Wikipedia pages on diverse landmarks, which have over 8,000 characters of contents to have abundant topics including history, design, tourism, and structures, and etc. For

[1] http://github.com/pkchat-focus/FoCus

[2] We gave the qualification test to the workers for a high-quality dataset and paid 166 qualified workers $5.5 for a single dialog.

	Knowledge Source	Persona Source	# Dialogs	# Average Turns	# Utterances
Wizard of Wikipedia (Dinan et al. 2018)	✓	✗	22,311	9.0	201,999
CMU-DoG (Zhou, Prabhumoye, and Black 2018)	✓	✗	4,112	31.0	130,000
PERSONA-CHAT (Zhang et al. 2018)	✗	✓	10,907	14.0	164,356
FoCus (Ours)	✓	✓	14,452	11.99	173,424

Table 1: Comparison of our FoCus dataset with other datasets. Our dataset is composed of 14,452 dialogues, which has 12 average turns, with 173,424 utterances. The utterances of FoCus dataset consider both knowledge and persona sources.

the persona sentences, we have 27,170 unique persona sentences related to landmarks' keywords implying its diversity. We provided a corresponding Wikipedia page as a knowledge source to the workers. To select out the pages with abundant descriptions about diverse aspects of the topic We only adopted the pages of which the number of the characters is over 8,000. The workers were instructed with two-step data creation procedure: **Step 1. Make a Persona** and **Step 2. Make a Dialog**.

Step 1. Make a Persona. In the FoCus dataset, we define `persona`, described by five sentences, as a personal background which can be any sentence about experience, preference, possession, hobby or interest. The workers were instructed to choose their own avatar and landmark. Then they make a virtual personal background regarding the landmark. To encourage the workers to generate topic-relevant persona, we let them to extract the keywords in the given Wikipedia page and make the persona sentences by means of the keywords. By creating persona based on the keywords, the topic and persona become closely related, which leads to more engaging dialog, as exemplified in Figure 2. Meanwhile, the workers were also allowed to create topic-agnostic persona sentences.

Step 2. Make a Dialog. After creating persona sentences, the workers were instructed to make a dialog by considering both persona and landmark knowledge. Unlike procedures done in previous datasets (Dinan et al. 2018; Zhang et al. 2018), they were instructed to make a multi-round dialog alone by alternating roles of *human* and *machine*, which enables more consistent and natural dialogs. We conducted the pilot study on the settings of creating dialog and concluded that the data from the single-person setup had high quality, especially in fusing persona and knowledge. As the person who asks the question knows better what knowledge one needs than the other person, the data from the single-person setup provided relevant and more customized answers.

To make customized and knowledgeable utterances, we gave the situation where the human asks a question regarding the landmark to the workers. In this situation, the machine answers the question by considering both `knowledge` and `persona` or only `knowledge`. As the human asks a question about the landmark which requires specialized knowledge to be answered, `persona`-only answer does not appear, which cannot give knowledgeable information to the user. For the first turn, we randomly gave one of the pre-generated questions so as to help the workers to smoothly start the first utterance of the dialog.

In addition, we also collected the grounding sources of machine's answers by letting the workers mark the sources they used, from `persona` or `knowledge`, when making answers. For instance, if they used `persona`, corresponding persona sentence was marked, and if they used Wikipedia `knowledge`, they indicate the referenced sentences in the Wikipedia page. These grounding information is used to evaluate the ability of models to ground the sources of their answers. The grounding abilities of the models can be quantitatively measured by proposed persona grounding (PG) and knowledge grounding (KG) sub-tasks, which will be described in the Experiments section.

Dataset Analysis

We report the comparison between our dataset and others with detailed statistics. In addition, characteristics of the customized answers in our dataset are analyzed.

Dataset Statistics We finally collected 14,452 dialogs with about 6 rounds per dialog on average. A comparison of our FoCus dataset with others is shown in Table 1, including the number of dialogs, average turns, utterances, and data sources used. We split the collected data into train, valid and test sets. The average length of the machine's utterances, which is about 141.13 in the train set, is much longer than that of the human's, which is about 40.94. It is because the machine provides the specialized knowledge when answering the question. Also, 44,518 of knowledge-only answers and 42,186 of persona-knowledge answer. The detailed statistics of our dataset are summarized in Table 2.

Types of Customized Answers The machine's answers can be categorized into three types according to their intent, i.e., *Inform*, *Confirm*, and *Suggest*. We describe the characteristics of each intent type. Note that Utt. stands for Utterances.

	Train	Valid	Test
# Dialogs	11,562	1,445	1,445
# Average Rounds	6.00	6.00	5.99
Avg. Length of *Human*'s Utt.	40.94	40.89	41.08
Avg. Length of *Machine*'s Utt.	141.13	145.42	146.67
# Knowledge-Only Answer	35,580	4,501	4,437
# Persona-Knowledge Answer	33,792	4,169	4,225
# Landmarks	5,082	1,305	1,299

Table 2: Statistics of FoCus dataset.

Inform. The answers that do not reflect the persona could be classified into *Inform*, which is similar to types of previous dialog datasets (Zhou, Prabhumoye, and Black 2018). This type of answers only utilizes the knowledge when making an answer. As exemplified in Figure 2, the answer that provides the size of the island is one of the examples.

Confirm. The intent of *Confirm* is to rephrase the user's persona and express the consent to it, as depicted in the first answer in Figure 2 . The answer of the machine confirms the user's preference for visiting Singapore. This type of answer is relatively more engaging than the answers with the *Inform* intention, as the given persona sentences are reflected. They are similar to the answers from Zhang et al. (2018). However, these answers still have a limited range, and the persona is not deeply utilized in the answers.

Suggest. Unlike above two types of answers, the answers with *Suggest* type recommends additional information that the users might like and enjoy or not suggest certain knowledge that users might hate or uncomfortable. This kind of answers give customized knowledge to the user by considering their persona, and they have not been introduced in other datasets. For example, the machine's answer that recommends the Universal Studios, because the user enjoys attractions during a tour, has the *Suggest* intention.

Model

We introduce the baseline models trained on our FoCus dataset, consisting of a *retrieval module* and a *dialog module*. The *retrieval module* retrieves the knowledge paragraphs related to a question, and the *dialog module* generates utterances of the machine by taking the retrieved knowledge paragraphs, human's persona, and previous utterances as inputs. An overview of our model is depicted in Figure 3.

Notation

The FoCus dataset is comprised of N dialogs and each dialog is composed of R rounds such that $D = \{(u_1^h, u_1^m), ..., (u_R^h, u_R^m)\}$ with the utterances of human u^h and machine u^m. The dialog is given with the corresponding persona and landmark knowledge. The human's persona is denoted as P, and knowledge documents about landmark are indicated as K. We further define the candidate sets of persona and knowledge, C_P and C_K, respectively, which are given at every turn and composed of the ground truth answers and distracting answers. Such candidates can be used to improve the grounding ability of agent by learning to select a ground truth answer among them, and more details are in Experiments section. The number of candidates of C_P and C_K are J and S, respectively.

Retrieval Module

To avoid excessive memory consumption, we present a retrieval module that enables narrowing the Wikipedia document down to five paragraphs K', which are related to the given utterance of human u^h. Among KNN (Fix and Hodges 1989), TF-IDF (Salton and Buckley 1988) and dense passage retrieval methods, we choose the TF-IDF score to retrieve the most related top 5 passages for the fast and efficient computation. To ensure its retrieval capability, BERTscore (Zhang et al. 2019a) is used to estimate how much the retrieved paragraphs are semantically similar to the gold knowledge. Note that the gold knowledge is reconstructed by the workers with their chosen sentences from the given Wikipedia paragraphs. The average BERTscore between the gold knowledge and the top 1 paragraph is about 83%, which is a relatively high score. As a result, TF-IDF score is used to choose five paragraphs, K', from the given knowledge document K which is utilized as the knowledge source for the answer, as shown in Figure 3. We calculate term frequency-inverse document frequency (TF-IDF) similarity score between the last question of human and possible knowledge paragraphs after the evaluation on the retrieved paragraph. The average token number of retrieved passages is about 132, and only the first 150 tokens are used as inputs.

Dialog Module

After selecting relevant knowledge paragraphs, the model first generate context-relevant representations to obtain the vectors that is highly relevant to the given knowledge, persona, and history. The representations are used to select the persona and knowledge from C_P and C_K, respectively. Chosen knowledge and personas are then concatenated with the dialogue history and then fed into the language modeling along with the machine's answer. Consequently, our training objectives are composed of language modeling for persona grounding, knowledge grounding and utterance generation among the given persona and knowledge candidates, C_P and C_K, which is trained in a multi-task learning (MTL) fashion (Ruder 2017; Zhang and Yang 2021). The number of candidates, J and S, are 5 and 10 respectively.

Context-Relevant Representation. Dialog module first makes a Context Relevant representation (CR) of the current dialog turn. Chosen knowledge paragraphs K' and a concatenation of persona and history $[P; U]$ are given as inputs. They are encoded by transformer, resulting $T(K')$ and $T([P; U])$ respectively, where T denotes a transformer model. Then, $T(K')$ is updated with the attention (Bahdanau, Cho, and Bengio 2014) mechanisms and concatenated with $T([P; U])$ resulting in the final representation CR.

Persona Grounding. To make a model that reflects the proper persona of the human when making answers, the model learns which persona to utilize, given the CR representation. As multiple persona sentences or none of them could be in the ground-truth answers, we train our model to discriminate each persona sentence to be used among the persona candidates. The special tokens are added to the each candidates. We utilize them by concatenating CR and the last hidden state representations of the special tokens from each candidate. The loss function is defined as follows:

$$L_{PG} = -\sum_{j=1}^{J} (q_j^* \log \text{Prob}([CR; h(p_j)]) + (1 - q_j^*) \log (1 - \text{Prob}([CR; h(p_j)]))), \quad (1)$$

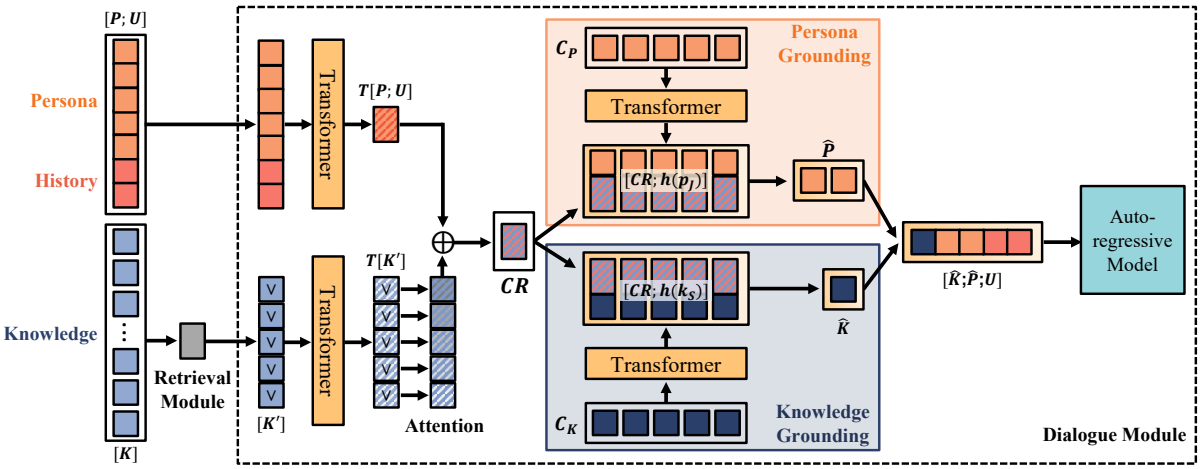

Figure 3: Overview of model architecture. The retrieval module selects five paragraphs K' from the documents of the given landmark. It goes through Transformers and is updated with attention mechanism. It is concatenated with the representation of Transformer-encoded sequence of persona and history, depicted as a cross in a circle. The CR is trained for the grounding tasks, and chosen persona and knowledge ($\hat{P}$ and $\hat{K}$) from the given candidate sets (C_P and C_K) are used to train the model's generation competence.

with q_j^* denoting a label defined as 1 if j-th persona sentence is ground-truth, 0 otherwise. $h(p_j)$ is the last hidden state representation of the special token of p_j. Prob($[CR; h(p_j)]$) is the estimated probability of the models.

Knowledge Grounding. The model also learns to use knowledge grounding to generate informative answers. The C_K consists of the ground-truth sentence and distracting candidates that are from the documents of different landmark. Given knowledge candidates at each round, the model is trained to choose one knowledge item that is expected to be used to answer the question by concatenating CR and the last hidden state representations of the special tokens from knowledge candidates:

$$L_{KG} = -\sum_{s=1}^{S} q_s^* \log \text{Prob}([CR; h(k_s)]), \tag{2}$$

with q_s^* denoting a label defined as 1 if s-th knowledge paragraph is ground-truth, 0 otherwise. $h(k_s)$ is the last hidden state representation of the special token of k_s. Prob($[CR; h(k_s)]$) is the estimated probability of the models.

Language Modeling. To build a generative agent, we model the machine's utterances in an auto-regressive manner. We consider two types of model structures, that are decoder-only and encoder-decoder. Following the previous works of Jelinek (1980); Bengio et al. (2003), the language modeling loss function is defined such that

$$L_{LM} = -\sum_{i=1}^{I} \log \text{Prob}(x_i | v, x_1, ..., x_{i-1}), \tag{3}$$

where Prob$(\cdot)$ denotes a probability of the langauge model, x_i is i-th token of u^m, I is the number of tokens and v stands for the sequence $[\hat{K}; \hat{P}; U]$ with concatenation of $\hat{K}$, $\hat{P}$, and U. $\hat{K}$ and $\hat{P}$ are the predicted candidates by the model in the knowledge grounding (KG) and persona grounding (PG) tasks respectively. Note that in the decoder-only model, $[\hat{K}; \hat{P}; U]$ are defined as the sequence of previous tokens, while they are used as the encoder inputs in the encoder-decoder model.

Full Objectives. The entire loss function aims to minimize the negative log-likelihood of language modeling and subtasks as in (Radford et al. 2019; Wolf et al. 2019). The full training objectives are defined as follows:

$$L = \lambda_{PG} L_{PG} + \lambda_{KG} L_{KG} + \lambda_{LM} L_{LM}, \tag{4}$$

where λ controls the proportion of each task during the training. In the experiments, λ_{LM}, λ_{PG}, and λ_{KG} were set to 10, 1 and 1, respectively. λ is chosen by the manual search.

Experiments

In this section we describe all the details of experiments including baselines, training settings and evaluation. We also analyze the experimental results and human evaluation of the dialog models trained on our dataset.

Language Model Baselines

We first describe the baseline language models, including transformer decoder, transformer encoder-decoder, GPT-2 and BART. By being trained with multi-tasks, those models are able to choose which persona and knowledge to use, as well as generate utterances. We implement the models based on the source code of HuggingFace's transformers (Wolf et al. 2020, 2019).

Models	Generation						Grounding (Acc.)	
	PPL	chrF++	BLEU	R-1	R-2	R-L	Persona	Knowledge
Decoder +PG +KG	228.69	0.1565	3.53	22.41	4.78	18.60	**67.83**	64.28
Enc-Dec +PG +KG	428.75	0.1345	2.79	18.45	2.81	14.80	**67.83**	64.52
GPT-2	17.42	0.1942	5.97	26.61	9.73	23.13	65.50	10.71
GPT-2 +PG	18.45	0.2221	5.63	25.56	9.12	22.20	**67.83**	9.25
GPT-2 +KG	**10.73**	0.2875	11.29	36.35	19.89	32.35	45.61	71.33
GPT-2 +PG +KG	11.45	0.2777	10.65	35.26	18.82	31.33	**67.83**	70.95
BART	26.55	0.1982	5.70	25.67	8.90	21.70	67.49	14.05
BART +PG	26.54	0.1932	5.36	25.35	8.43	21.40	**67.83**	14.75
BART +KG	15.84	**0.2946**	**11.64**	**36.19**	**19.90**	**31.84**	53.78	**73.00**
BART +PG +KG	23.25	0.2887	11.28	35.35	19.12	31.06	**67.83**	71.70

Table 3: Experimental results of the baseline models on the test set. The models are evaluated by generation metrics, including perplexity (PPL), chrF++, SacreBLEU, ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L), and accuracy for persona grounding task and knowledge grounding task.

Transformer-based Models. We train the models with a transformer (Vaswani et al. 2017) structure. Both decoder-only model and encoder-decoder model are used to generate the utterances. To evaluate the effectiveness of pre-training, we set transformer layers to have the same structure with the following pre-trained language models.

Pre-trained Language Models. We adopt GPT-2 (Radford et al. 2019) and BART (Lewis et al. 2019) as pre-trained decoder-only and pre-trained encoder-decoder models, respectively, which are known to show remarkable performances in language generation by training a colossal number of parameters on a massive corpus.

Experimental Setup

We train GPT-2_{Small}, which has 12 layers and 12 attention heads with 768 embedding dimensions, and BART$_{Base}$, which has 6 layers each in both the encoder and decoder, and 12 attention heads with 768 embedding dimensions. We use a batch size of 4 with a gradient accumulation of 32. Adam optimizer is used, and the learning rate is set as 6.25e-5, where β_1 = 0.9, β_2 = 0.999 with linear decay. For the hyperparameter settings, we adopt the initial hyperparameters from the models trained on PERSONA-CHAT (Zhang et al. 2018) and Wizard-of-Wikipedia (Dinan et al. 2018) datasets. Among the candidates, we choose the hyperparameters that showed the best performance. Fine-tuning them on the entire data with 2 epochs takes approximately 10 hours with one RTX-8000 GPU. For the utterance generation, we use the nucleus sampling with top-p = 0.9 and sampling temperature with 0.7. The maximum sequence length is set to 20. Generation and grounding evaluation takes about 30 minutes.

Automatic Score Evaluation

To evaluate model's ability to give fluent, attractive and informative utterances, sub-tasks for measuring the ability of generating customized responses (*generation*) and discriminating which source to reflect (*grounding*) are provided.

Task 1 - Generation. To evaluate the generation competence, the perplexity (PPL) is used to measure the fluency as in other generation tasks (Zhang et al. 2018; Dinan et al. 2018). The chrF++ (Popović 2017) score, SacreBLEU(Post 2018), and recall-oriented understudy for gisting evaluation (ROUGE-1-F, ROUGE-2-F, ROUGE-L-F) (Lin 2004) are adopted to assess how close the generated answer is to the original answer.

Task 2 - Grounding. In addition, we evaluate the models' grounding abilities by our proposed PG and KG tasks, which enable us to test whether the models choose the proper persona and knowledge among the given candidates to generate an answer. As an answer of the machine that utilizes different persona and knowledge at each turn, we provide the persona candidates and knowledge candidates for every round. Whereas C_P consists of five given persona sentences, C_K includes the ground-truth sentences of Wikipedia and distracting candidates that have the same number of sentences from the other documents on different landmarks. We measure the accuracy of persona grounding and knowledge grounding persona selection and knowledge selection respectively.

Analysis. As shown in Table 3, we experiment with transformer-based decoder model, encoder-decoder model, GPT-2 and BART. We analyze their generation abilities on the test set. Out of the transformer-based models, the decoder-only model shows higher generation performance than the

Model	Rank	Fluency	Engagement	Consistency
Human	**1.05** (0.31)	**4.15** (1.54)	**4.08** (1.53)	**4.06** (1.47)
GPT-2	2.64 (0.48)	2.85 (0.93)	2.95 (0.98)	2.76 (0.99)
BART	2.31 (0.52)	3.13 (1.14)	3.18 (1.08)	3.10 (1.04)

Table 4: Human evaluation. The models trained with PG and KG are evaluated their utterances compared to the gold data of human. The value in the parenthesis indicates standard deviation of the scores.

Persona
1. I live in a building 2. I find heritage-listed buildings interesting 3. I am from Australia 4. I have never been to Queensland 5. I wish to visit Queensland
Landmark
https://en.wikipedia.org/wiki/Thorps_Building
Dialog
Human : Cool! What is it? *(1) BART* : The is a beautiful urban burial ground that contains a collection of highly intact funerary monuments and furniture dating *(2) BART +KG* : This is Thorps Building, a heritage-listed commercial building. *(3) BART +PG* : It is a historic burial ground located in Australia where you are from. *(4) BART +PG +KG* : It is a heritage-listed commercial building, you may have heard of it since you are a fan

Table 5: An example of conversations between human and BART_{Base}. (1) generates fluent utterance, but it is not closely related to the given persona and knowledge. (2) makes an informative answers, and (3) generates more user-aware answer. (4) seems to generate the most plausible utterance by fusing both persona and knowledge.

encoder-decoder model. In the grounding task, they show comparable performances. GPT-2_{Small} and BART_{Base} models are adopted as pre-trained language models, and they are trained to generate the machine's utterances. To investigate the effectiveness of the grounding task, we additionally train the models with or without two grounding sub-tasks. In the generation task, the language models trained with knowledge grounding (KG) task show high scores, especially BART trained with KG is the highest on the most of generation scores. However, their persona grounding (PG) accuracy is lower than others, which means that they are not good at choosing proper persona for each turn. The language models trained with both PG and KG show slightly lower but comparable performances in the generation task and, but they show competent scores in both of two grounding sub-tasks. Since all the results are rounded to two decimal point, numbers from PG seem to be the same. The best results of PG are converged to a certain number and it indicates the upper bound of the baseline models. Also, our experimental results indicate that the high automatic score on the generation task does not always guarantee the high grounding ability. The experiments suggest the need of versatile generative models that are able to not only make fluent utterances, but also select proper sources and fuse them competently.

Human Evaluation

To evaluate the fluency, engagement, and consistency in the utterances of machine on a numerical scale from 1 to 5, we randomly selected 20 dialogues generated by the models which are in the test set. We set up three questions and specified the level of answers with likert scale (Likert 1932). In addition, we asked human evaluators, the MTurk workers [3], to rank the each examples in order of which conversation shows the most human-like utterances by the machine following Cho and May (2020). Rank is scaled from 1 to 3 and the lower number indicates the better quality. The survey results are shown in Table 4. The gold data made by human shows the best scores on all criteria of fluency, engagement and consistency, which ranks first. Among GPT-2 and BART, note that they are trained on PK and KG, BART is shown to outperform GPT-2 on the all criteria. The result shows that the quality of the gold data surpasses the models' generation. In spite of the pre-trained models' massive parameters and their abilities, their responses, given the context, are much less engaging, fluent and consistent than those of humans which means that our dataset is considerably challenging.

Answer Type	Well-grounded utterances (%)
Knowledge-only	98.94
Knowledge-Persona	94.52

Table 6: Grounding quality assessment. The numbers indicate the proportions of the well-grounded utterances with knowledge, and both knowledge and persona respectively.

Grounding Quality Assessment

With the human evaluators, we evaluate the grounding quality of the dataset. We asked the workers to assess whether the answers in each utterance included Wikipedia knowledge or both Wikipedia knowledge and persona sentences. We had each dialogs evaluated by five independent workers [4] with the randomly selected 200 dialogs in our dataset. The results in Table 6 shows the proportions of well grounded utterances. The proportions of well-grounded utterances with knowledge-only are about 99% and and those of knowledge-persona grounded answers are over 94%.

Related Work

To build dialog agents that can interact with people in multi-turn conversations, several datasets have been introduced (Ritter, Cherry, and Dolan 2010; Danescu-Niculescu-Mizil and Lee 2011; Lowe et al. 2015; Wu et al. 2016; Li et al. 2017; Mostafazadeh et al. 2017; Shuster et al. 2018; Fan et al. 2019). Despite these datasets, the dialog agents merely answer the question without considering the user or specialized knowledge.

To generate customized answers to the users, attempts have been made to endow the agent with the user's emotion, preference, and experience (Rashkin et al. 2018; Shuster et al. 2018; Urbanek et al. 2019; Boyd et al. 2020). Zhang et al. (2018) introduces a dataset that includes each speaker's preference and experience, where persona sentences describing two speakers are given. Because both speakers are only provided with persona sentences, one speaker simply confirms

[3] We paid 4 qualified workers $2 for a single evaluation on the dialog.

[4] We paid 20 qualified MTurk workers $2 for a single evaluation on the dialog.

what the other speaker likes or dislikes in the dialog. Even though agents generate answers that react or express sympathy, they hardly give a document-grounded answer that fits the user's preference and experience.

While the user-centered dialog datasets have appeared, datasets and agents that aim to improve the level of knowledge in the answer with additional documents has been in parallel released (Dinan et al. 2018; Zhou, Prabhumoye, and Black 2018; Moghe et al. 2018; Qin et al. 2019; Gopalakrishnan et al. 2019; Cho and May 2020; Zhou et al. 2020; Santhanam et al. 2020). Dinan et al. (2018) is a dialog dataset where the agent retrieves the Wikipedia pages on diverse topics and generates responses to the questions. Although these data have a concept of persona, they do not contain customized answers to the listener. Similar to Dinan et al. (2018), Zhou, Prabhumoye, and Black (2018) introduces a document-grounded dataset that includes specified documents from Wikipedia articles about popular movies. These datasets mainly consist of answering the question without considering the user's information, and it leads to excessive and needless answers.

There have been efforts to blend several datasets (Shuster et al. 2019; Smith et al. 2020) to build an intelligent agent which has multiple abilities learned from various datasets. Despite the previous datasets, the capability of machines to respond in a dialog is still insufficient, compared to that of humans. Specifically, to answer a question, retrieving the knowledge while considering the user's background information is beyond current dialog agent's abilities.

Conclusion

In this work, we have introduced the FoCus dataset that contains the customized responses by utilizing both persona and the Wikipedia knowledge. To validate the effectiveness of our dataset, we adopted and trained the language models on the FoCus dataset. Along with the generation tasks, we evaluate the grounding abilities of the models with provided PG, and KG sub-tasks. The experiments demonstrated that the pre-trained models show high performance on the generation task, but it does not necessarily lead to high grounding performance, and may limit in the grounding abilities. As shown in human evaluation and grounding quality assessment, our dataset is proven to be natural but complicated for the machines to mimic. We believe our FoCus dataset can contribute to build more human-like agents which gives customized answers with proper knowledge. We will also additionally annotate the type of the intents for each answer to let the models learn the purpose during generating answers. In the future, the models trained with our dataset can be utilized in the situation where the specialized knowledge is required depending on the user's persona in the form of personal assistants. We hope that the researches aim to make dialog agents more attractive and knowledgeable with grounding abilities to be explored.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). Also, this work was supported by NCSOFT NLP Center.

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Janvin, C. 2003. A neural probabilistic language model. *The journal of machine learning research*, 3: 1137–1155.

Boyd, A.; Puri, R.; Shoeybi, M.; Patwary, M.; and Catanzaro, B. 2020. Large scale multi-actor generative dialog modeling. *arXiv preprint arXiv:2005.06114*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cho, H.; and May, J. 2020. Grounding Conversations with Improvised Dialogues. *arXiv preprint arXiv:2004.09544*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Danescu-Niculescu-Mizil, C.; and Lee, L. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Fan, A.; Jernite, Y.; Perez, E.; Grangier, D.; Weston, J.; and Auli, M. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Fix, E.; and Hodges, J. L. 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3): 238–247.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; Hakkani-Tür, D.; and AI, A. A. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH*, 1891–1895.

Jelinek, F. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980*.

Joshi, C. K.; Mi, F.; and Faltings, B. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Majumder, B. P.; Jhamtani, H.; Berg-Kirkpatrick, T.; and McAuley, J. 2020. Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions. *arXiv preprint arXiv:2010.03205*.

Moghe, N.; Arora, S.; Banerjee, S.; and Khapra, M. M. 2018. Towards exploiting background knowledge for building conversation systems. *arXiv preprint arXiv:1809.08205*.

Mostafazadeh, N.; Brockett, C.; Dolan, B.; Galley, M.; Gao, J.; Spithourakis, G. P.; and Vanderwende, L. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Popović, M. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, 612–618.

Post, M. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*.

Qin, L.; Galley, M.; Brockett, C.; Liu, X.; Gao, X.; Dolan, B.; Choi, Y.; and Gao, J. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. *arXiv preprint arXiv:1906.02738*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 172–180.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Salton, G.; and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5): 513–523.

Santhanam, S.; Ping, W.; Puri, R.; Shoeybi, M.; Patwary, M.; and Catanzaro, B. 2020. Local Knowledge Powered Conversational Agents. *arXiv preprint arXiv:2010.10150*.

Shuster, K.; Humeau, S.; Bordes, A.; and Weston, J. 2018. Image Chat: Engaging Grounded Conversations. *arXiv preprint arXiv:1811.00945*.

Shuster, K.; Ju, D.; Roller, S.; Dinan, E.; Boureau, Y.-L.; and Weston, J. 2019. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *arXiv preprint arXiv:1911.03768*.

Smith, E. M.; Williamson, M.; Shuster, K.; Weston, J.; and Boureau, Y.-L. 2020. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. *arXiv preprint arXiv:2004.08449*.

Urbanek, J.; Fan, A.; Karamcheti, S.; Jain, S.; Humeau, S.; Dinan, E.; Rocktäschel, T.; Kiela, D.; Szlam, A.; and Weston, J. 2019. Learning to speak and act in a fantasy text adventure game. *arXiv preprint arXiv:1903.03094*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Weyand, T.; Araujo, A.; Cao, B.; and Sim, J. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2575–2584.

Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

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

Wu, B.; Li, M.; Wang, Z.; Chen, Y.; Wong, D.; Feng, Q.; Huang, J.; and Wang, B. 2019. Guiding Variational Response Generator to Exploit Persona. *arXiv preprint arXiv:1911.02390*.

Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; and Li, Z. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Xu, M.; Li, P.; Yang, H.; Ren, P.; Ren, Z.; Chen, Z.; and Ma, J. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation. *arXiv preprint arXiv:2002.02153*.

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, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

Zhang, Y.; and Yang, Q. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*.

Zhou, H.; Zheng, C.; Huang, K.; Huang, M.; and Zhu, X. 2020. KdConv: a Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *arXiv preprint arXiv:2004.04100*.

Zhou, K.; Prabhumoye, S.; and Black, A. W. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.