Personalized Response Generation via Generative Split Memory Network

Yuwei Wu[♣]*, **Xuezhe Ma**[♠], **Diyi Yang**[♢]
[♣] Shanghai Jiao Tong University, will8821@sjtu.edu.cn
[♠] University of Southern California, xuezhema@usc.edu
[♢]Georgia Institute of Technology, dyang888@gatech.edu

Abstract

Despite the impressive successes of generation and dialogue systems, how to endow a text generation system with particular personality traits to deliver more personalized responses remains under-investigated. In this work, we look at how to generate personalized responses for questions on Reddit by utilizing personalized user profiles and posting histories. Specifically, we release an open-domain *single-turn* dialog dataset made up of 1.5M conversation pairs together with 300k profiles of users and related comments. We then propose a memory network to generate personalized responses in dialogue that utilizes a novel mechanism of splitting memories: one for user profile meta attributes and the other for user-generated information like comment histories. Experimental results show the quantitative and qualitative improvements of our simple split memory network model over the state-of-the-art response generation baselines. The dataset and code are available here.

1 Introduction

Building human-like conversational systems, in particular chit-chat agents, has been a long-standing goal in language technology communities. Unlike task-oriented dialog agents that focus on completing specific tasks (Wen et al., 2017; Eric et al., 2017; Lei et al., 2018; Lowe et al., 2015), chit-chat agents need to dynamically interact with people, understand the meaning of human conversations (Hovy and Yang, 2021), and thereby make better responses to improve user experience.

Despite the recent successes on building chit-chat agents using data-driven approaches (Ritter et al., 2011; Banchs and Li, 2012; Serban et al., 2016; Li et al., 2016c; Parthasarathi and Pineau, 2018), lack of a consistent personality is still one of the common issues. The main reason is that these models are often trained over conversations spoken by different people, ignoring their personality (Li et al., 2016b; Wei et al., 2019; Zhang et al., 2018). As shown in Table 1, different people responded differently to the same input question due to their diverse background including basic personal information and attitudes towards different things. Thus, it becomes essential to incorporate **personalization** into the modeling and evaluation of response generation and eventually chit-chat agents.

There have been several personality-related dialogue datasets built for evaluating models' performances in personalized conversations, such as PERSONA-CHAT dataset (Zhang et al., 2018) and Facebook's Reddit dataset (Mazare et al., 2018). The PERSONA-CHAT dataset was collected by intentionally assigning annotators to predefined personas described by a set of sentences instead of their real personality. Such artificially generated conversations cannot adequately represent respondents and their personalities which would lead to

Question: Where do you live and what is something you are doing today?
Responses:
A: I live in *Mongolia* and I will be making some good *sandwiches* today.
B: Midwest America, I will be skyping my *brothers* and going to *band practice* today.
Question: What's your "go to" when you're sad?
Responses:
A: I *listen to horror stories* for some reason.
B: I love to *read* or *listen to sad music*.
Respondent Profile:
A: *Gender:* female; *Favorites: sandwich*; *Possessions:* Russian class; *Residence: Mongolia*; *Asia*;
B: *Family: brothers*; *Self-description: guitarist*; *Favorite: fakebooks*; *Residence: America*;
Respondent Comment Histories:
A: I often fall asleep while *listening to horror stories*.
B: *Listening to sad music*, I know it adds fuel to fire but the flame will burn out quicker and you'll feel better soon.

Table 1: Example conversation pairs with respondents' profile and posting histories, with related information from profile and histories in *blue* and *red* respectively.

*The work is mainly done when YW was a visiting student at Georgia Institute of Technology.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1956–1970
June 6–11, 2021.

dataset bias problems. For example, an introvert annotator can hardly imitate the response of a person with sociable personas. Moreover, the number of personas covered by this corpus is limited.

Today's social media platforms such as Reddit and Twitter provide us with good opportunities to build a large scale of collections of naturally occurring conversations (Xifra and Grau, 2010; De Choudhury and De, 2014; Schrading et al., 2015) and also make it possible to provide consistent personalities. For instance, Facebook's Reddit dataset represents each user by a set of sentences chosen from their comment histories heuristically. However, they also acknowledged that these persona sentences might not well represent a general trait of users due to the limitation of their heuristic rules for sentence retrieval (Mazare et al., 2018).

In this work, we introduce a personalized Reddit dataset **PER-CHAT**, an open-domain response generation dataset consisting of 1.5M conversations and 300k users. PER-CHAT covers finer-grained personal information for users, including discrete user attributes such as gender, residence, self-description and favorites inferred based on users' self-reported messages on Reddit, and contextual information such as their comments (§3). Based on PER-CHAT, we propose a simple **generative split memory network** to incorporate diverse personal information, with a novel mechanism of splitting memories: one memory representation for user meta attributes (e.g., profile) and the other for user activity information (e.g., comment histories), respectively (§4). Experimental results show that our generative split memory network outperforms state-of-the-art response generation baselines both quantitatively and qualitatively (§5).

2 Related Work

Personalized Generation Datasets Much attention has been paid to construct personalized dialog datasets. Built upon the bAbI dialog dataset, Joshi et al. (2017) extended it to include information such as gender, age and dietary preference. This domain-specific dataset was then used to train goal-oriented dialog models for several restaurant reservation tasks. There are also several dialog datasets that focus on chit-chat scenarios, such as PERSONA-CHAT dataset (Zhang et al., 2018), Reddit dataset (Al-Rfou et al., 2016), Twitter dataset (Li et al., 2016b) and PersonalDialog dataset (Zheng et al., 2020). PERSONA-CHAT (**PC**) dataset consists of 1k different personas, and annotators are asked to conduct conversations according to assigned personas. The Reddit dataset and Twitter dataset simply use user ID information without any specific user information to indicate personalization. The PersonalDialog dataset (**PD**) (Zheng et al., 2020), collected from a Chinese social media Weibo, contains three kinds of personality traits ("gender", "location", "age") for each user. On the other hand, Mazare et al. (2018) introduced personalization from Reddit (**PCR**) by incorporating the persona of each user with a (randomly chosen) subset of his/her posting comments. Zhong et al. (2020) further extended their datasets with annotated empathy information (**PEC**). In this work, we combine those two different ways of gathering personalization signals of users, i.e., meta profile attributes and users' posting histories, and provide a more comprehensive, large scale personalized dataset derived from natural social conversations.

Personalized Generation Models Current dialog models can be divided into ranking-based models and generation-based models. Ranking-based models (Al-Rfou et al., 2016; Mazare et al., 2018; Zhang et al., 2018) focus more on the task of response selection that is to pick the best response from a pool of random candidates. In contrast, generation-based models attempt to generate response directly from any given input questions. Under personalized dialog settings, Zhang et al. (2018) claimed that ranking-based models performed better than generative models on their personalized dataset, suggesting that building personalized generation models are more challenging.

With the development of recent large scale social media data and the success of sequence to sequence framework (Serban et al., 2016; Shang et al., 2015; Sutskever et al., 2014), several personalized response generation models have been proposed, and we can only mention a few here due to space limits. Li et al. (2016b) introduced the Speaker Model and the Speaker-Addressee Model that encoded user-id information into an additional vector and fed it into the decoder to capture the identity of the speakers. Kottur et al. (2017) further extended these speaker models into multi-turn conversations. In addition to using user id to capture personal information, Zhang et al. (2018) proposed a profile memory network that utilizes a memory network for encoding persona sentences. To further utilize personal traits, Zheng et al. (2020) proposed

an attention mechanism to incorporate these user-related attributes in the decoding stage. Recently, there are a few works using meta-learning and reinforcement learning to enhance mutual persona perception Madotto et al. (2019); Kim et al. (2020); Majumder et al. (2020). However, few models have taken into account different potential sources of personalization signals such as profile attributes and comments. Our work conducts persona-aware representation learning by combining these two sources. Note that our split memories architecture is similar to Joshi et al. (2017) , but differs in tasks and memorizing histories. In our work, we focused on memorizing relevant history comments instead of dialog histories in multi-turn chat settings.

Evaluation Metrics Most response generation models utilize perplexity, BLEU (Papineni et al., 2002) and recently BERTScore (Zhang et al., 2019) and Moverscore (Zhao et al., 2019) for evaluation (Serban et al., 2016; Xing et al., 2018). For evaluating personalization, Zheng et al. (2020) proposed to measure the accuracy of predicting personality traits by firstly training classifiers for different personality traits such as gender and age. However, for certain trait categories such as hobbies and location, it is quite difficult to train a reliable classifier. In terms of evaluating persona consistency between generated sentences and given user comments, Madotto et al. proposed consistency score using NLI models pre-trained on Dialog NLI dataset (Welleck et al., 2019), which is a corpus based on Persona dataset, with NLI annotation between persona description sentences and dialogues utterance. In this paper, we introduce an automatic metric for evaluating persona consistency between user profiles and these generated sentences.

3 Dataset Construction

This section describes how we construct an open-domain single-turn dialog dataset with personalization information from Reddit, together with dataset analysis[1]. Specifically, we used r/AskReddit[2], one of the most active subreddits based on an online subreddit ranking system sorted by number of active users[3]. Users on r/AskReddit are encouraged to write clear and direct questions, and most posted questions are about open-ended discussion on a variety of topics, without definite or correct answers or professional knowledge, making r/AskReddit a suitable place to model personalization in open domain dialogue systems.

Data Preprocessing. We collected all submissions under r/AskReddit as questions and their subsequent comments as responses. Each submission and one of its direct comment form a (question, response) pair in our corpus, i.e., single-turn dialogues. Furthermore, we stripped away potential markdown and Html syntax tokens and replaced all forms of url links, emails, and digits in our corpus with unique tokens "url", "email" and "digit" respectively. We also processed replicated words and punctuation to their standard form via a set of regular expressions, e.g., "cooooool" is converted into "cool" and "!!!!!" to "!".

Vocabulary and Conversation Pairs. We use a vocabulary of 50,257 entries the same as Dialogpt (Zhang et al., 2020), since they pretrained their models using the full Reddit data. To avoid lengthy questions or responses, we pruned the conversation pairs based on the statistics (see Figure 3 in Appendix A). Questions that exceed 100 words and responses with over 40 words are excluded. In total, there are 1,566,653 conversation pairs.

3.1 Personalization Information

To augment our dataset with personalization information, we collected three sources of user-related information: (1) *user IDs* which are unique usernames for their Reddit accounts; (2) *comment histories*, which are all the comments a user has posted on Reddit; (3) *user profile attributes* such as gender, residence, favorites and etc.. To collect these user-specific information, we first filtered out inactive users — a user who has made less than 100 comments during the recent year. There remain 301,243 users after removing inactive users.

User Comment Histories. Users' comment histories can often signal their personal preferences toward topics or even texting habits as shown in Table 1, thus it is beneficial to collect these histories. We obtained a user's comment histories by querying the Pushshift Reddit API[4]. Since (1) it is infeasible for models to operate on the scale of thousands of comments and (2) applying persona extraction process rather than randomly picking up comments can improve model's performance

[1] Similar process can be employed to our raw data to obtain MULTI-TURN dialog datasets.

[2] https://www.reddit.com/r/AskReddit/

[3] https://www.topsubreddits.com/

[4] https://github.com/pushshift/api

in personalization suggested by Mazare et al., we designed an information retrieval (IR) system to automatically pick up query-related comment histories for each user. Specifically, We utilized semantic embedding based similarity between each query and a comment to obtain a smaller set of candidates $\mathbb{M}$, following similar retrieval mechanisms as Ritter et al. (2011); Wang et al. (2013). That is, given the input question, we retrieve top l comments that have the highest cosine similarity scores with the query to construct the user's comment histories. The embedding used in IR systems is the averaged contextual embeddings from pretrained BERT-large models(Devlin et al., 2019). *Respondent Comment Histories* of Table 1 shows some example query-related histories we extracted from a user's comments.

User Profile. The persona extraction process to construct comment histories might lose valuable user's attributes such as their residence, favorites which are also helpful in generating personalized responses. To this end, we further conduct a finer-grained entity extraction mechanism over all of user's past histories. User profile information was viewed as entities extracted from histories using similar methods as the popular Reddit user analysis site SnoopSnoo[5]. Following the categories provided by the site, we first divided user attributes into eight types, including *"pets"*, *"family"*, *"residence"*, *"favorites"*, *"partner"*, *"possessions"*, *"gender"*, *"self-description"*, where 'possessions" refers to personal possessions owned by users such as users' guitars; "favorites" means users' favorite items and people mentioned by the user and "self-description" denotes concepts that users use to describe themselves such as their occupations. We then applied different extraction regular expressions for different categories. For example, we would gather a noun as "favorites" if it is found after "*like,love,..*" in certain comments.

Examples for these attributes are shown in *Respondent Profile* of Table 1. Unlike some social media platforms such as Weibo, users on Reddit do not provide very specific profile information. Thus, we need to extract these entities based on their histories, and also check the reliability of such profile information. We manually checked whether such extracted user attributes actually corresponded to users' comments via a small corpus study (details in Appendix B), and found that in over 85% cases, our entity extracting process is quite reliable for capturing users' basic information.

[5] https://github.com/orionmelt/snoopsnoo

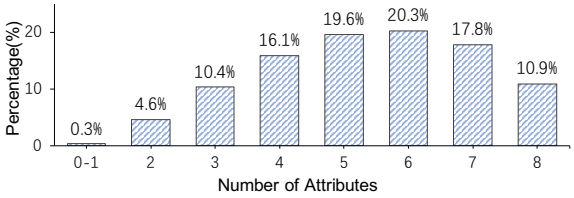

Figure 1: Distribution of users' number of attributes.

Attr	Pets	Family	Residence	Favorites
Percent(%)	29.1	70.9	39.1	62.7
Attr	Partner	Possession	Gender	Self-description
Percent(%)	39.1	99.7	99.5	82.8

Table 2: Coverage rate for each attribute.

	Relevant	**Irrelevant**	**Unsure**
Numbers	434	63	3
Percentage	86.8%	12.6%	0.6%

Table 3: Query-response relevance annotation.

	Train	**Dev**	**Test**
# Queries	439996	5523	5559
# Response	1528218	19224	19211

Table 4: Statistics of train, dev, and test set.

User Profile Analysis. We conducted in-depth analyses to show the coverage rate of each attribute out of the eight profile attributes in our collected corpus, as described in Table 2. We found that gender and possession have very high coverage rates above 99%, and other attributes have different coverage rates, ranging from 29.1% to 82.8%. Since it is unnecessary that users contain value under every attribute type, we also computed the percentage of users who have the corresponding number of attribute types. Figure 1 showed that most users have around 4 to 7 attributes.

Question-Response Relevance To examine the quality of our constructed corpus, especially the question-response relevance, we randomly sampled 500 question-response pairs from our corpus, and asked for annotators from Amazon Mechanical Turk to rate them. Each pair is judged by three raters on whether a response appropriately responded to the given question. Raters can select from 'Yes', 'No', and 'Unsure' if they are unsure about the relevance. We obtained an intra-class correlation coefficient of 0.63 , indicating good

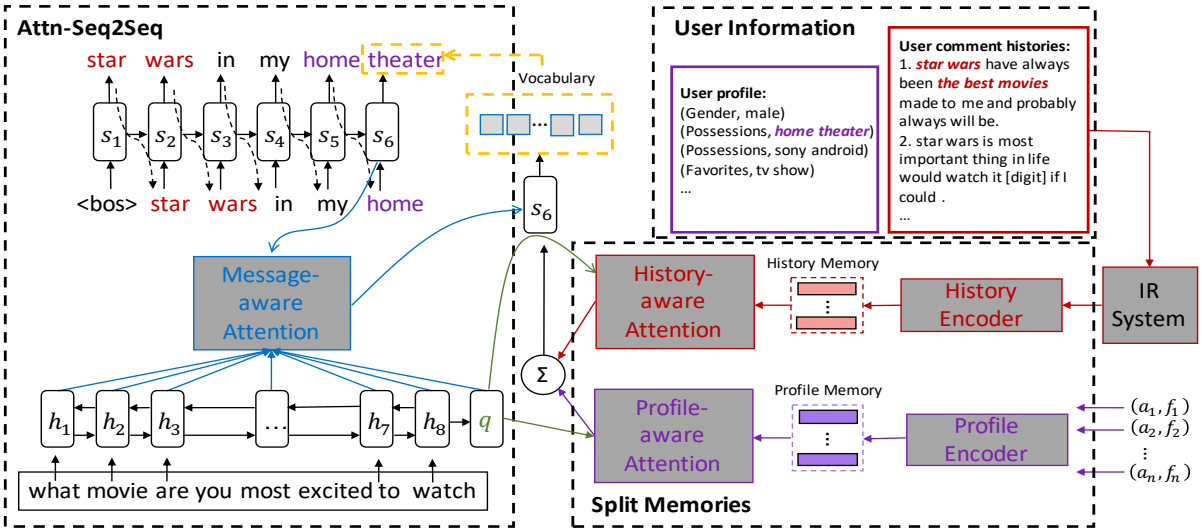

Figure 2: The overall diagram of Generative Split Memory Network.

Dataset	Source	Comments	Profile	Size	Public
PC	Crowd-Sourced	Yes	No	151K	Yes
PCR	Reddit	Yes	No	700M	No
PEC	Reddit	Yes	No	355K	Yes
PD	Weibo	No	Yes	20.83M	No
PER-CHAT	Reddit	Yes	Yes	1.5M	Yes

Table 5: Comparisons between PER-CHAT and related datasets. PC denotes PERSONA-CHAT (Zhang et al., 2018). PCR denotes the persona-based dialog datasets from Reddit (Mazare et al., 2018). PEC denotes persona-based empathetic conversation (Zhong et al., 2020). PD denotes the PersonalDialog dataset (Zheng et al., 2020). The size denotes the number of expanded conversations.

annotation agreement(Cicchetti, 1994). We categorized a pair as relevant if the majority of annotators vote 'Yes'. As summarized in Table 3, 86.8% of the pairs are found to be relevant, suggesting a reasonable quality of our corpus.

Most questions in our corpus has around with 2 to 3 responses. We randomly sampled 1% of questions and their corresponding responses as the development set, 1% as the testing set, and the remaining 98% as the training set. Table 4 summarizes the detailed statistics.

Comparisons with Related Datasets Table 5 shows the comparisons between our datasets and the related ones. The biggest advantage of our dataset is that it has both comment histories and user profiles while being 5-10 times bigger than any prior publicly available dataset. In our following experiments in section 5, we show the necessity to provide both dimensions of personalized information. In terms of comments, we applied pre-trained IR system to extract query related comments instead of simply rule-based filtering used in datasets such as PCR and PEC. In terms of user profiles, we provide more diverse categories with eight main types, larger than PD datasets which only contains age, gender and location. By utilizing social media data, our dataset allows for more diverse personality and natural dialog patterns with over 300k users than datasets collected by human (e.g. PC).

4 Generative Split Memory Network

This section presents our generative models for personalized response generation, which generate responses conditioned on given questions and respondents' personal information[6].

Let a conversation C be a tuple of Question, Response and respondent (User) $C := (Q, R, U)$. A user $U = (ID, P, M)$ consists of three sources of information — username (ID), profile attributes $P = (f_1, f_2, \ldots, f_n)$, and user's comments $M = (m^1, m^2, \ldots)$ where f_i is given as a key-value pair $f_i =< k_i, v_i >$ and M is a set of comment histories the user made.

To better incorporate personal information of different dimensions, we propose a generative memory network with split memories for user profile and user comment history, respectively. The intuition lies in that interpersonal meta attributes and comment patterns may influence the respondents' responses differentially. The overall model architecture is shown in Figure 2. Our model is built

[6]Multi-turn dialogue generation can be considered in our framework by encoding additional contexts in memories.

upon a standard seq2seq model with attention. For a given conversation C, we first feed the input comments M through retrieval system to get query-related comment set $\mathbb{M}$, and the memory network encoder computes the representations of related comment history. In parallel, profile attributes are added as separate profile memory by another encoder. At each time step, the decoder utilizes the aggregated representations of comment histories and profile attributes to generate the final response.

4.1 User Profile and Comment Histories

For a user U with the profile P, we view P as the user's attribute sequence and employ a shared word embedding in encoder to encode the attribute key sequence as $\boldsymbol{e_k} = (\boldsymbol{e_{k_1}}, \boldsymbol{e_{k_2}}, ..., \boldsymbol{e_{k_n}})$ and to encode each entry in attribute value sequence as $\boldsymbol{e_v} = (\boldsymbol{e_{v_1}}, \boldsymbol{e_{v_2}}, ..., \boldsymbol{e_{v_n}})$ respectively. The final set of the user profile representation $\boldsymbol{H_p}$ is defined as $\{\boldsymbol{e_{k_1}} \odot \boldsymbol{e_{v_1}}, ..., \boldsymbol{e_{k_n}} \odot \boldsymbol{e_{v_n}}\}$, which is considered as the profile memory.

For encoding comment sentences, we encoded the retrieved comments $\mathbb{M}$ for a user U as individual memory representations in a memory network, similar to Zhang et al. (2018). Instead of applying weight functions to word vectors of each entry, we feed the comments to the RNN encoder to get the set of encoded history memories denoted as $\boldsymbol{H_m}$.

4.2 Split Memories

We then pass $\boldsymbol{H_p}$ and $\boldsymbol{H_m}$ to a split memory encoder. The memory network separately attends to the encoded split memories with given query vector $\boldsymbol{q}$ over K hops as follows:

$$a_p^k = Softmax\left(\boldsymbol{H_p} \cdot W_1 \cdot w_p^k\right) \quad (1)$$

$$w_p^{k+1} = (a_p^k)^T \cdot \boldsymbol{H_p} + w_p^k \quad (2)$$

$$a_m^k = Softmax\left(\boldsymbol{H_m} \cdot W_2 \cdot w_m^k\right) \quad (3)$$

$$w_m^{k+1} = (a_m^k)^T \cdot \boldsymbol{H_m} + w_m^k \quad (4)$$

where $W_1, W_2 \in \mathbb{R}^{d \times d}$ and $w_p^1 = w_m^1 = \boldsymbol{q}$. The outputs from both memories w_p^K and w_m^K are summed to get the representation $\boldsymbol{O^K}$ and is then fed into decoder side.

The memory decoder utilizes the memory network and RNN. The RNN decoder takes as input the previous hidden state and previous target word embedding and generates the hidden state $\boldsymbol{h_t}$ at the time step t. The vocabulary distribution P_{vocab} for time step t is generated as follows:

$$P_{vocab} = Softmax(W_3[\boldsymbol{h_t}; \boldsymbol{O^K}]) \quad (5)$$

where $W_3 \in \mathbb{R}^{|V| \times 2d}$ is a trainable parameter.

5 Experiment

5.1 Implementation Details

Our implementation is based on the Pytorch version of OpenNMT (Klein et al., 2017)[7]. We used the pre-trained Dialogpt word embedding (Zhang et al., 2020). The hidden size of the encoder and decoder were set to 1024. The embedding size is the same as the memory size and the RNN hidden size. We used AdamW (Loshchilov and Hutter, 2018) as our optimizer with an initial learning rate of 5e-5 and a linear decay learning rate schedule. The dropout rate was set to 0.1. The batch size was selected in $\{16, 32, 64, 128\}$. The maximum number of iteration steps was set as 20000 with an early stop if no improvement over perplexity on dev set. To generate hypothesis sentences, we used nucleus (top-p) filtering (Holtzman et al., 2019) without any re-scoring techniques. The cumulative probability for top-p filtering is set as 0.4.

5.2 Baseline Models

We introduced several baselines to compare with our generative split memory network (**GSMN**).

- **Attention-Seq2Seq:** a standard seq2seq model with attention mechanisms proposed by Luong et al. (2015), without utilizing any personal information.
- **Speaker Model:** Similar to (Li et al., 2016b), we employed an additional vector to model the respondent A.
- **Generative Memory Network w/ History:** Following Zhang et al. (2018), we encoded the retrieved comments as individual memory representation in a memory network to incorporate comment histories $\mathbb{M}$.
- **Generative Memory Network w/ Profile:** We designed a memory network model to incorporate user profiles P by doing attention over user attributes (Zheng et al., 2020) .
- **Dialogpt:** The state-of-the-art large-scale pre-trained response generation model on 147M Reddit corpus(Zhang et al., 2020).

[7] https://github.com/OpenNMT/OpenNMT-py

Model	User Information	Perplexity	BLEU	C	PC-Score
Attention-Seq2Seq	-	110.920	1.324	0.300	0.00989
Speaker Model	username	92.607	1.329	0.301	0.0102
Generative Memory Network	profile	75.635	1.592	0.304	0.0131
Generative Memory Network	history	80.000	1.664	0.305	0.0120
Dialogpt	-	58.723	3.246	0.306	0.0182
Dialogpt	profile+history	36.764	5.894	0.309	0.0237
Generative Split Memory Network	profile+history	72.173	1.700	0.306	0.0152
Dialogpt w/ Split Memories	profile+history	**33.519**	**7.047**	**0.311**	**0.0337**

Table 6: Automatic results on PER-CHAT test sets.

- **Dialogpt w/ Split Memories** We directly combined the split memory network with the pretrained models , i.e. we applied the same architecture as GSMN in the decoder side and used pre-trained Dialogpt as the encoder.

5.3 Evaluation Metrics

We evaluated the baselines and our generative split memory network using several widely-used metrics, including perplexity, BLEU, and BERTScore, to compare models' performances in generating appropriate responses. **Perplexity** is used to measure how the outputs fit test data (Vinyals and Le, 2015; Serban et al., 2016). Models with lower perplexity scores are found to demonstrate better performance to generate grammatical and fluent responses (Xie et al., 2019; Zheng et al., 2020). We also used **BLEU** (Papineni et al., 2002; Li et al., 2016a; Galley et al., 2015) with n-grams (n=1) to measure how many n-grams in generated responses overlap with those in reference responses. To evaluate persona consistency between user comments and generated sentences, Madotto et al. proposed consistency **C** score using sequence classification model trained on Dialog NLI dataset (Welleck et al., 2019), a corpus based on Persona dataset, with NLI annotation. For given comments p_js and generated sentence u, the consistency score is given as follow:

$$\mathbf{NLI}\left(u, p_j\right) = \begin{cases} 1 & \text{if } u \text{ entails } p_j \\ 0 & \text{if } u \text{ is independent to } p_j \\ -1 & \text{if } u \text{ contradicts } p_j \end{cases}$$

$$\boldsymbol{C}(u) = \sum_{j}^{m} \mathbf{NLI}\left(u, p_j\right) \tag{6}$$

Note that models with higher consistency C scores tend to generate more persona consistent responses with user's comments. In our settings, m is set 10 for max number of given comments.

PC-Score In addition to the aforementioned evaluation metrics, we also designed a metric called Profile Consistency Score (**PC-Score**) to measure a model's performance in generating persona consistent responses with given user profiles. The idea is similar to the *entity score* in knowledge enhanced conversation tasks (Zhou et al., 2018), which computes the number of entities for each response and aims to measure the model's ability to select the concepts from the commonsense knowledge. Instead of calculating the number of entities selected from knowledge base per response, we did a micro-average over the number of entities selected from the profile of each user to capture personalization.

Manual Evaluation We conduct manual annotations to examine the consistency of those models. Here, the consistency refers to that the generated responses should be consistent for the same user when similar questions are asked. For example, when asked "*Where are you from?*" and "*Where is your hometown?*", the generated responses should be consistent in certain granularity for the same user. To this end, we randomly chose ten users from our user population set and designed 20 questions. Half of these questions are related to basic personal information, e.g., residence and gender, and the other half is related to personal interests and attitudes such as favorite activities. Detailed experiments are shown in the Appendix D.

We generated 200 responses from each model, and asked annotators on Amazon Mechanical Turk to judge such consistency based on two criteria: (1) *model consistency*: whether or not a given generated sentence is consistent under the same group of questions; (2) *personalization consistency*: whether or not a given generated sentence is consistent with a user's personal information. Raters were asked to rate between 1 to 3 for model consistency, where 1 means "*Not consistent at all*", 2 means "*Slightly*

Model	User Information	Consistency	Personalization Consistency
Attention-Seq2Seq	-	1.40^{**}	0.29^{**}
Speaker Model	username	1.88^{**}	0.31^{*}
Generative Memory Network	history	2.08^{*}	0.35^{*}
Generative Split Memory Network	profile+history	**2.29**	**0.43**

Table 7: Manual evaluation results. Here, *Personalization Consistency* can take value from -1 to +1. $*$ indicates significant difference with the best result (t-test, p-value<0.05); and $**$: p<0.01.

consistent", 3 means "*Very consistent*". For personalization consistency, raters rate whether the generated sentences match either the provided user profile attributes or user comments[8]. Note that when annotating personalization consistency, the turkers were not able to see the user ids; all the user information shown to them is publicly available on Reddit to protect user information.

5.4 Results

Table 6 summarizes different evaluation metrics on test set. We found that the **Speaker Model** boosted the **Attention-Seq2Seq** baseline with a decrease of 18.3 in perplexity, 3.14% increase in PC-Score, similar to Li et al. (2016b). However, because the user set is quite large, the performance improvement of Speaker model is limited. The generative memory network that incorporates either user profile or comment history demonstrates improvement compared with Attention-Seq2Seq in terms of all the metrics. Either generative memory network outperformed the speaker model, suggesting that network with additional user information has a better ability to generate semantic consistent and personalized responses. Furthermore, our proposed network (**GSMN**) significantly outperforms all other baselines, with a decrease of 38.7 in perplexity and a 28.4% increase in BLEU over Attention-Seq2Seq. By applying the split memories, **Dialogpt w/ Split Memories** further outperform **Dialogpt** in terms of all metrics. It shows that current pre-trained dialog models lack the ability of personalized memorization though they were found to be effective in memorization and generalization on a wide range of classical dialog tasks(Zhang et al. (2020)). This further justifies the necessity of our datasets.

In terms of human evaluations for both consistency and personalization consistency (Table 7), our network also demonstrates consistent improvement over the baselines. Note that we do not apply any copy mechanisms in our models, the generation of personalized entities is purely depending on representation learning. It shows Split memory network outperforms the baselines on generating sentences with better personalization and consistency. To further examine the effectiveness of our generative split architecture, we also compared with Dialogpt that utilizes both profile and comment history to generating responses. In this setting, we included both profile and history as individual memory and did attention mechanism over this memory, shown as **Dialogpt + profile + history**. We observed that our generative split memory network still achieved better performance.

Overall, this shows that incorporating both dimensions of personalization information, i.e., user profile and user comment history, can boost models' performances for response generation and split architecture for generation is better at utilizing these two different personalization signals.

5.5 Discussion

Diverse Responses Conditioned on Users: Table 12 in Appendix C shows some example responses generated by our GSMN, together with the Seq2Seq baseline. The examples are randomly sampled from our test set. Given different user profiles, GSMN is more effective and faithful to the profile attributes of different users in generating user-specific responses. For example, our model can identify user's profiles like families and gender when being asked about the most reliable person while the baseline's answer is more like consensus and applicable for any user. More examples are in Table 14 in Appendix C.

Consistency Analysis: Example outputs from baselines and our model are described in Table 13 in Appendix C. The Seq2Seq model was a bit inconsistent in answering the same group of questions. The speaker model showed consistency to some degree since the answers "California" and "Florida" are quite close in the word embedding space, but failed due to the lack of user information.

[8]Comments that have highest similarity scores with the queries are used as references.

Compared with these baselines, GSMN is much better at generating both consistent and personalized responses. For example, when asked favorite activities, GSMN responds consistently and is also sensitive to personalized information since "my dog" identifies the pet attribute of the respondent.

6 Conclusion

In this paper, we introduced a large-scale open-domain personalized dataset PER-CHAT and proposed a generative split memory network to utilize both user profile information and commenting histories for the task of response generation. Experimental results showed that our proposed model significantly outperformed several state-of-art baseline models, both quantitatively and qualitatively. Future research could build upon our work on single-turn response generation to further model personalization in multi-turn conversations.

7 Ethical Considerations

For the annotation, each worker on Amazon Mechanical Turk was paid 0.1$ per selection task (matching the United States federal minimum wage). To ensure quality, we chose only master crowd-workers who had more than 5000 HITs approved and with an approval rate larger than 95%.

Considering the privacy violation problems our dataset may bring about, we followed Reddit's term of use for user content—based on Reddit API Terms of Use, users are granted with license to display the user content through application[9].

We have taken careful procedures to protect users' privacy concerns. First, our introduced PER-CHAT dataset will be shared for **academic use only**. We only released raw data from pushshift.io(Baumgartner et al. (2020)), and open-sourced our scripts for preprocessing user attributes and models for reproducibility. Note that, the user attributes used in our work are identified based on users' self-reported statements via regular expressions matching. This research study has been approved by the Institutional Review Board (IRB) at the researchers' institution.

Generation models trained on Reddit sometimes tend to generate toxic or inappropriate responses as pointed out by Dialogpt (Zhang et al., 2020). Due to this reason, we followed their best practices to deal with released version of decoding scripts[10].

[9] 2.d. in Reddit API Terms of Use

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

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*.

Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*.

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

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and William B Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *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)*, pages 445–450.

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

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *NAACL*.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Satwik Kottur, Xiaoyu Wang, and Vítor Carvalho. 2017. Exploring personalized neural conversational models. In *IJCAI*, pages 3728–3734.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

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

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

I Sutskever, O Vinyals, and QV Le. 2014. Sequence to sequence learning with neural networks. *Advances in NIPS*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2019. Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7290–7294. IEEE.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

TH Wen, D Vandyke, N Mrkšíc, M Gašíc, LM Rojas-Barahona, PH Su, S Ultes, and S Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*, volume 1, pages 438–449.

Yubo Xie, Ekaterina Svikhnushina, and Pearl Pu. 2019. A multi-turn emotionally engaging dialog model. *arXiv preprint arXiv:1908.07816*.

Jordi Xifra and Francesc Grau. 2010. Nanoblogging pr: The discourse on public relations in twitter. *Public Relations Review*, 36(2):171–174.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2020. Personalized dialogue generation with diversified traits.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

A Conversation Pairs and Response Distribution

We examined the distribution for query-response pairs and the statistic result is shown in the Figure 3. Questions that exceed 100 words and responses that are longer than 40 words are excluded. This led to 88% of the original pairs.

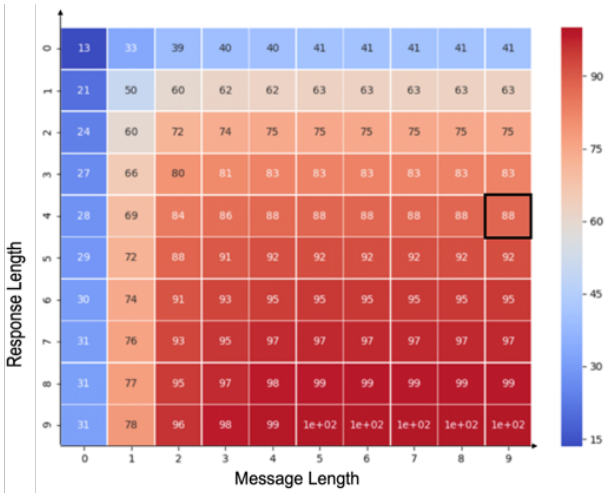

Figure 3: Message and response length distribution.

Figure 4 illustrates the distribution of the number of responses under same question in our dataset. We found that questions with 2 to 3 responses are the majority.

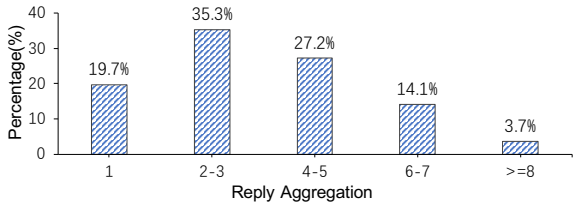

Figure 4: The average number of replies per question.

B Reliability of API-based User Attribute Information

To examine the reliability of the extracted information, we conducted human annotation validations. Specifically, we randomly selected 50 users from our population set together with their attributes. And we shared the user attributes and source comments with annotators and asked them to judge whether the user attributes corresponded to the comments. If the source comment truly reflects user's corresponding attributes, they should give label "Right", otherwise "Not Right". For attribute types with more than one value, such as possessions, "Right" means all the values are truly related and "Partly right" means some are related and some are not and "Not Right" means all the values are not related.

Attribute	Pets	Family	Residence	Favorites
Percentage	28	70.9	46	56
Attribute	Partner	Possession	Gender	Self-description
Percentage	38	98	100	86

Table 8: Coverage rate for each attribute in those sampled 50 users.

Attribute	Pets	Family	Residence	Favorites
Right / %	85.7	82.4	82.6	96.4
Partly Right / %	14.3	8.8	0	0
Not Right / %	0	8.8	17.4	3.6
Attribute	Partner	Possession	Gender	Self-description
Right / %	100	83.7	86.0	65.1
Partly Right / %	0	12.2	0	18.6
Not Right / %	0	4.1	14.0	16.3

Table 9: Attribute reliability annotation results.

Table 8 shows the coverage rate for each attribute in selected users. The distribution aligns well with the overall coverage rate and shows that the sampled users are representative. Table 9 shows the reliability annotation result. The percentage for each attribute in annotation result is calculated among users with value in that attribute. As shown in Table 9 all of the attributes show a high reliability rate by considering "Right" and "Partly Right". Although all the user attributes are inferred from what user has said about himself/herself, there still exists information that does not represent his/her personal attributes. Table 10 shows examples of positive and negative label results for some attributes.

Source Comments	**Consistency?**
Gender:	
I am a thin girl that has trouble...	Right
If I was a girl who was...	Not Right
Favorites:	
I LOVE Halloween and delight in...	Right
I like the force.	Not Right
Residence:	
I live in San Diego county.	Right
I live just outside of Boston.	Not Right
Possession:	
I have it on my pandora playlist.	Right
I wanted to start up my own prison.	Not Right
Self-description:	
I'm a rapper who...	Right
I was basically a zombie.	Not Right
Family:	
I asked my mother if she loved me...	Right
I met our mother-a documentary...	Not Right

Table 10: Annotation examples for different attribute types with attributes in red.

C Diverse Responses for Models

Table 12 and Table 14 show additional responses generated by different models on our dataset. With profile and related comments, our model can generate not only user-attribute related entity but also is capable of capturing different users' attitudes towards other people or things. For example, the responses generated by GSMN properly reflected the user's positive or negative opinions when Trump was mentioned.

D Consistency Experiments

In Table 11 and 15, we show all 20 questions we designed and profiles of four sampled users in the manual evaluation experiments. From question no.1 to question no.10, the questions are related to basic personal information and from question no.11 to question no.20, the questions are related to personal attitudes. The answer of every two adjacent questions should be considered consistent. Table 16 shows examples of top related comments for given questions and given user.

Questions for basic information:
1. Where are you from ?
2. Where do you come from ?
3. Do you have a girlfriend or boyfriend ?
4. Are you single ?
5. What do you do for living ?
6. What is your job ?
7. Are you male or female ?
8. What is your gender ?
9. What pet do you have ?
10. Do you have any pets ?
Questions for personal attitudes:
11. Who is your most reliable one?
12. Who do you rely on most?
13. What is your most upset experience?
14. What upsets you most?
15. What is your attitude toward Trump?
16. Will you vote for Trump?
17. Who is your favourite person?
18. Who do you love most?
19. What do you like to do most?
20. What do you enjoy most?

Table 11: Questions in human evaluation experiments.

Question: Who is the one person you could rely on if you were in a real bad situation?
Seq2Seq: My best friend.
User A: *Gender: male*; *Family: wife* **Gold:** Wife or guy I used to do things with a long time ago. **GSMN:** My wife. She's my wife.
User B: *Gender: female*; *Family: husband* **Gold:** My dad. Guy would melt the ice caps if I asked him to. **GSMN:** My husband. He's a good guy.
User C: *Gender: male*; *Partner: girlfriend* **Gold:** My girlfriend. **GSMN:** My ex. She' s my best friend.

Table 12: Example responses generated by our GSMN.

User A: *Residence:Canada*
Question: Where are you from ?
Seq2Seq: I' m in the UK. **Speaker:** I' m in California. **GMN:** New York city. **GSMN:** I' m from Canada.
Question: Where do you come from?
Seq2Seq: I' m from US **Speaker:** I' m in Florida. **GMN:** I' m in the US. **GSMN:** I' m from Canada. It's pretty <unk>.
User B: *Comments:I love dogs, I really do.*
Question: What do you like to do most ?
Seq2Seq: I like to watch movies. **Speaker:** I like to play video games. It' s so much fun. **GMN:** I' ve always wanted to be able to play video games. **GSMN:** I like to play with my dogs. I love them.
Question: What do you enjoy most?
Seq2Seq: I like to play video games. **Speaker:** I like to go to the gym. **GMN:** Being in a good mood. **GSMN:** I like to play with my dogs.

Table 13: Examples from consistency evaluation.

Question: Writing a christmas card to Mr. Trump! What should it say?
Seq2Seq: I'm going to buy a new one. #
User A: **Gold:** You' re a piece of garbage and we all hate you so much. **GSMN:** I'm not a Trump supporter.
User B: **Gold:** Thank you for doing a great job. **GSMN:** It's a wonderful life.
Question: Gamers of Reddit what, is the most overrated game that critics have shilled for?
Seq2Seq: I' m not sure if it' s a game or not, but I' d say <unk>. #
User A: **Gold:** Most AAA games. **GSMN:** The last of us. It' s so boring.
User B: **Gold:** Fortnite, and other battle royales. **GSMN:** Fortnite, it' s a bad game.
Question: What was the last thing you laughed at?
Seq2Seq: I' ve seen a lot of funny jokes.
User A: **Gold:** A comment on reddit about a guy. **GSMN:** I laughed at a meme.
User B: **Gold:** My best friend' s toddler dipping his cookie in ketchup. **GSMN:** My friend' s birthday.
User C: **Gold:** My cat headbutting me in the face. **GSMN:** My cat' s laugh.
Question: You have one movie to watch for the rest of your life , what is it ?
Seq2Seq: I' d like to see a movie called "<unk>". #
User A: **Gold:** Star Wars The Force Awakens. **GSMN:** I' m a fan of Star Wars.
User B: **Gold:** Prestige [digit] watches. **GSMN:** Requiem for a Dream, <unk>.

Table 14: Example responses generated by GSMN and baseline. # indicates poor-quality response.

User1:
Gender: male Residence: Canada Pets: dog Family: sister; father; mother Partner girlfriend Favorites: ice cream Self-description: good artist; newbie Possessions: team rocket hoodie; video games; junk food addiction
User2:
Gender: female Residence: Germany Pets: cat Family: sister; father; mother Partner: husband Favorites: honey whine; buckwheat; beef Self-description: christian bit; smartest person Possessions: university subject chemistry
User3:
Gender: male Residence: Illinois Pets: cat; dog Family: father; mother Partner: wife Favorites: pumpkin cheesecake; chili dogs; Self-description: native American man Possessions: American accent; hearing loss; church family; food aversion; stomach problem
User4:
Gender: male Pets: dog Family: father Partner: girlfriend Favorites: adventure

Table 15: Sampled user profiles in human evaluation experiments.

Question: What upsets you most?
User1: People who pronounce my name wrong.
User2: When people around me show self restraint it reminds me of my failures and this feels bad.
User3: It really hurts when someone says something.
User4: Anything unhealthy. Also, anything that makes someone feel sad.
Question: Who is your favorite person?
User1: My sister and my dog.
User2: My husband who is my best friend and my biggest supporter.
User3: Most popular girls in school. #
User4: My guitar. And my lamp. And my girlfriend. #
Question: What do you like to do most?
User1: I love food more than people . Though if my dog wanted some of my food, I' d love him enough to share it.
User2: The most magical thing for me is and has always been winter solstice. The rebirth of the sun. I always bake a sweet sun-bread.
User3: I love dogs, I really do.
User4: I like adventure time. Feels like old school cn.

Table 16: Top retrieved comments for given questions in human evaluation experiments. # indicates that the retrieved top comments are not well related to our designed questions.