VALUENET: A New Dataset for Human Value Driven Dialogue System

Liang Qiu[1], Yizhou Zhao[1], Jinchao Li[2], Pan Lu[1], Baolin Peng[2], Jianfeng Gao[2], Song-Chun Zhu[1]

[1]UCLA Center for Vision, Cognition, Learning, and Autonomy
[2]Microsoft Research, Redmond
liangqiu@ucla.edu

Abstract

Building a socially intelligent agent involves many challenges, one of which is to teach the agent to speak guided by its value like a human. However, value-driven chatbots are still understudied in the area of dialogue systems. Most existing datasets focus on commonsense reasoning or social norm modeling. In this work, we present a new large-scale human value dataset called VALUENET, which contains human attitudes on 21,374 text scenarios. The dataset is organized in ten dimensions that conform to the basic human value theory in intercultural research. We further develop a Transformer-based value regression model on VALUENET to learn the utility distribution. Comprehensive empirical results show that the learned value model could benefit a wide range of dialogue tasks. For example, by teaching a generative agent with reinforcement learning and the rewards from the value model, our method attains state-of-the-art performance on the personalized dialog generation dataset: PERSONA-CHAT. With values as additional features, existing emotion recognition models enable capturing rich human emotions in the context, which further improves the empathetic response generation performance in the EMPATHETICDIALOGUES dataset. To the best of our knowledge, VALUENET is the first large-scale text dataset for human value modeling, and we are the first one trying to incorporate a value model into emotionally intelligent dialogue systems. The dataset is available at https://liang-qiu.github.io/ValueNet/.

Introduction

Value refers to desirable goals in human life. They guide the selection or evaluation of actions, policies, people, and events. A person's value priority or hierarchy profoundly affects his or her attitudes, beliefs, and traits, making it one core component of personality (Schwartz 2012). In dialogue systems, modeling human values is a critical step towards building socially intelligent chatbots (Qiu et al. 2021b). By considering values, we can estimate user behavior and cognitive patterns from their utterances and generate responses that conform to the robot's persona configuration. For example, the robot is set to be aware of human values, and it invites Jerry to drink beers, but Jerry replies, "*You know that is tempting but is not good for our fitness*". The bot could read from the dialogue that Jerry prefers a healthy and self-disciplined lifestyle and steer its recommendation to healthier options in the future.

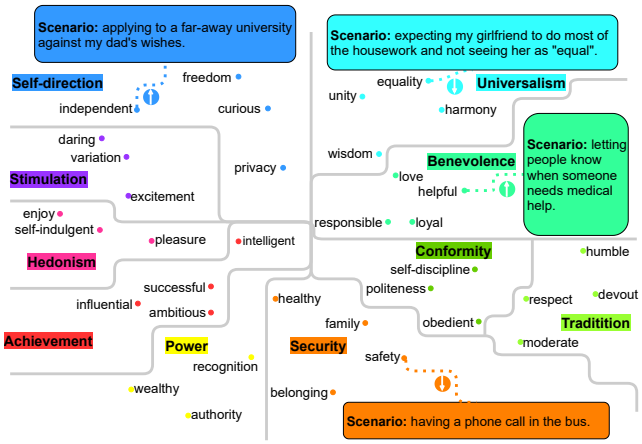

Figure 1: The presented VALUENET dataset with curated social scenarios organized by Schwartz values (Schwartz 2012).

The development of socially intelligent chatbots has been one of the longest-running goals in artificial intelligence. Early dialogue systems such as Eliza (Weizenbaum 1966), Parry (Colby, Weber, and Hilf 1971), and more recent SimSimi[1], Panda Ichiro (Okuda and Shoda 2018), Replika (Fedorenko, Smetanin, and Rodichev 2018), XiaoIce (Zhou et al. 2020), were designed to mimic human behavior and incorporate emotional quotients (EQ) to some extent. There are also datasets and benchmarks for studying related problems, such as emotion recognition (McKeown et al. 2010; Hsu et al. 2018; Poria et al. 2019; Ghosal et al. 2020), personalized dialogue generation (Zhang et al. 2018; Liu et al. 2020), and empathetic dialogue generation (Rashkin et al. 2019). Even though value plays a fundamental and critical role in human EQ, there is a lack of explicit modeling of values in the dialogue domain, based on social domain theory. We have seen recent efforts about crowdsourcing social commonsense knowledge base or benchmarks (Forbes et al. 2020; Sap et al. 2019; Lourie, Bras, and Choi 2021;

[1]https://simsimi.com/

Hendrycks et al. 2020; Hwang et al. 2021; Gabriel et al. 2021). However, it is not clearly shown how an agent can leverage this knowledge to estimate the users' value priorities or guide its own speaking and actions. In this paper, we aim to alleviate this problem and investigate the usage of a learned value function.

We start the study by curating a knowledge base of human values called VALUENET. Samples with value-related scenarios were identified based on value-defined keyword searching. Next, we asked Amazon Mechanical Turk workers about how the provided scenarios will affect one's value. This is based on the assumption that values underlie our attitudes; they are the guideline by which we evaluate things. Workers assess behaviors/events positively if they promote or protect the attainment of the goals we value. Behaviors/events are evaluated negatively if they hinder or threaten the attainment of these valued goals. The whole process gives us a large-scale (over 21k samples) multi-dimensional knowledge base of value. Figure 1 shows the overall structure of VALUENET. Each split represents a value dimension identified in the theory of basic human values (Schwartz 2012). The figure also illustrates the value-related keywords and scenarios. The circular arrangement of the values represents a motivational continuum. By organizing data in such a structure, we anticipate the VALUENET to provide comprehensive coverage of different aspects of human values.

Next, we develop a Transformer-based value model to evaluate the utility score suggesting the positive or negative judgment given an utterance. We provide a detailed analysis of learning with multiple Transformer variants. Then we conduct a wide range of experiments to demonstrate that the value model could benefit EQ-related dialogue tasks: *(i)* By finetuning a generative agent with reinforcement learning and the reward from our value model, the method achieves state-of-the-art performance on the personalized dialogue dataset: PERSONA-CHAT (Zhang et al. 2018); *(ii)* By incorporating values as additional features, in EMPATHETICDIALOGUES (Rashkin et al. 2019), we improve the emotion classification accuracy of existing models, which further facilitates the empathetic response generation; *(iii)* Visualization of the value model shows that it provides a numerical way of user profile modeling from their utterances.

In all, our contributions are two-fold. First, we present a large-scale dataset VALUENET for the modeling of human values that are well-defined in intercultural research. Second, we initiate to develop the value model learned from VALUENET to several EQ-related tasks and demonstrate its usage for building a value-driven dialogue system. Our methodology can be generalized to a wide range of interactive situations in socially aware dialogue systems (Zhao, Romero, and Rudnicky 2018), and human-robot interactions (Yuan and Li 2017; Liang, He, and Anthony'Chen 2021).

Related Work

An abundance of related work inspires our work. Our work aims to make contributions to dialogue systems by incorporating the theory of human value. The dataset we collect shares a similar nature with multiple social commonsense benchmarks and knowledge bases. Besides, we apply our VALUENET for various dialogue tasks related to EQ.

Theory of Human Value and Utility

In the field of intercultural research, Schwartz (2012) developed the theory of basic human values. The theory identifies ten basic personal values that are recognized across cultures and explains where they come from, as shown in Figure 1. The closer any two values in either direction around the circle, the more similar their underlying motivations are; the more distant, the more antagonistic their motivations. Note that dividing the value item domain into ten distinct values is an arbitrary convenience. It is reasonable to partition the value items into more or less fine-tuned distinct values according to the needs and objectives of one's analysis[2]. Similarly, in the economics field, the concept of utility (Fishburn 1970) is initially defined as a measure of pleasure or satisfaction in economics and ethics that drives human activities at all levels. Therefore, when we teach agents to speak and act in a socially intelligent way, an approach considering human value utilities should be adopted. In this paper, we aim to learn a utility function for each dimension of value and steer the dialogue system response generation accordingly.

Social Commonsense Benchmarks

Hendrycks et al. (2020) present the ETHICS dataset, a benchmark that assesses a language model's knowledge of basic concepts of morality. SCRUPLES (Lourie, Bras, and Choi 2021) is a large-scale dataset with ethical judgments over real-life anecdotes, motivated by descriptive ethics. SOCIAL-CHEM-101 presented by Forbes et al. (2020) is a corpus that catalogs rules-of-thumb as basic concept units for studying people's everyday social norms and moral judgments. They also propose Neural Norm Transformer to reason about previously unseen situations, generating relevant social rules-of-thumb. SOCIAL IQA (Sap et al. 2019) is a large-scale benchmark for commonsense reasoning about social situations. He et al. (2017) present a task and corpus for predicting the preferable options from two sentences describing the scenarios that may involve social and cultural situations. Instead, in this work, we release a new dataset VALUENET that provides annotation of human attitudes from different value aspects.

Emotionally Intelligent Dialogue Datasets

Several datasets are presented to study emotion dynamics in dialogues. DailyDialog (Li et al. 2017) is a multi-turn dialogue dataset, which reflects the way of daily communication and provides emotion labels for speakers. Hsu et al. (2018) present EmotionLines with emotions labeling on all utterances in each dialogue based on their textual content. MELD (Poria et al. 2019) is an extension of EmotionLines for multi-modal multi-party emotion recognition. McKeown et al. (2010) record a corpus SEMAINE of emotion-

[2]A refinement of the theory (Schwartz et al. 2012), partitions the same continuum into 19 more narrowly defined values that permit more precise explanation and prediction. We use the original 10-dimension version for simplicity in this paper.

ally coloured conversations. Ghosal et al. (2020) propose a framework COSMIC for emotion recognition in conversations by considering mental states, events, actions, and cause-effect relations. DialogRE (Yu et al. 2020) is the first human-annotated dialogue-based dataset for social relation inference (Qiu et al. 2021a). PERSONA-CHAT (Zhang et al. 2018) (revised in ConvAI2 (Dinan et al. 2020)) provides natural language profiles of speakers. Based on PERSONA-CHAT, Liu et al. (2020) propose a transmitter-receiver-based framework with explicitly human understanding modeling to enhance the quality of personalized dialogue generation. EMPATHETICDIALOGUES (Rashkin et al. 2019) is a dataset that provides 25k conversations grounded in emotional situations. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion.

The VALUENET Dataset

During decision-making, people tend to pick the choice that aligns more with their own values. This work aims to provide a transferable knowledge base for human value modeling in natural language. To collect the VALUENET dataset, we curated social scenarios with value-related keywords and further annotated them via Amazon Mechanical Turk. Each sample in VALUENET is a social scenario description labeled with the annotator's attitude through a specific value lens.

The entire dataset is organized in a circular structure as shown in Figure 1, aligning with the theory of basic human values (Schwartz 2012). The theory identifies ten universal values that are recognized throughout major cultures. The circular structure reflects the dynamic relations among these values, *i.e.,* the pursuit of some value may result in either accordance with another value or a conflict with another value. The ten distinct values can be further organized into four higher-order groups.

- **Openness to change**: self-direction, stimulation
- **Self-enhancement**: hedonism, achievement, power
- **Conservation**: security, conformity, tradition
- **Self-transcendence**: benevolence, universalism

We describe the collection details of the VALUENET in the following sections.

Social Scenario Curation

We curated a set of 21,374 social scenarios from the large-scale social-related database SOCIAL-CHEM-101 (Forbes et al. 2020). Value-related scenarios are retrieved with value keywords after lemmatization and stemming. There are three sets of keywords identified for each dimension of Schwartz value: (1) the keywords in the original definition of each value in Schwartz's paper (Schwartz 2012); (2) words that share a similar meaning, words that are often used to describe the original keywords, and words that are triggered by (strongly associated with) the original keywords[3]; (3) words that are near the original keywords in the GloVe (Pennington, Socher, and Manning 2014) embedding space. The value keywords are verified and confirmed by humans as listed in Figure 2.

[3]We use datamuse (https://www.datamuse.com/api/) for this purpose.

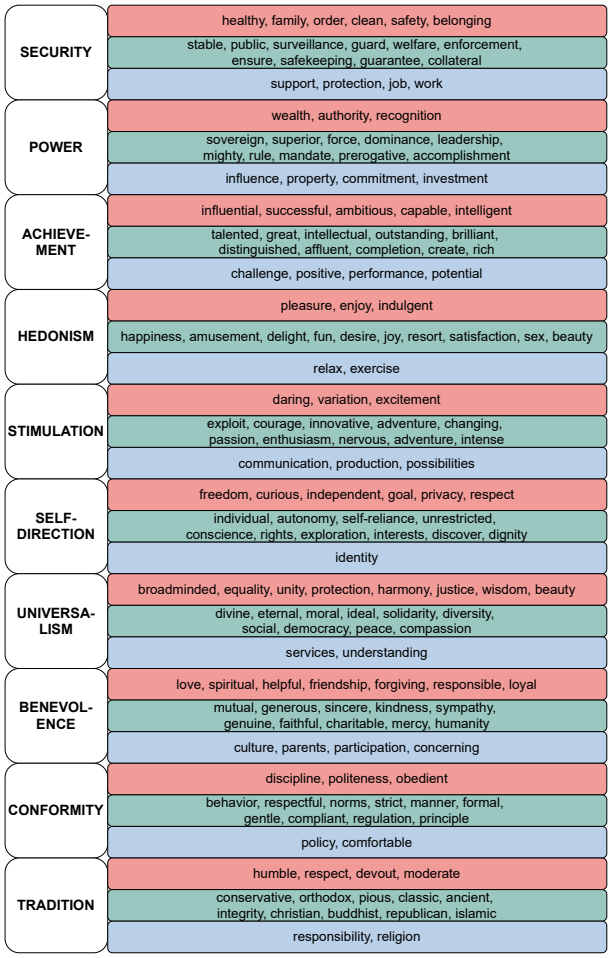

Figure 2: Ten universal human values and related keywords for social scenario curation. **Red**: keywords in the original value definition (Schwartz 2012); **Green**: associated keywords found with datamuse; **Blue**: associated keywords found with GloVe embedding.

Value-Aspect Attitude Annotation

We crowdsourced people's attitudes to the curated scenarios on Amazon Mechanical Turk (AMT). Figure 3 shows an example.

We follow a strict procedure to select qualified workers and ensure the workers understand the concept of each value we ask. In Figure 3, the definition of BENEVOLENCE is shown to the workers throughout the entire annotation process. To further help the understanding, we include three examples in each assignment with correct answers being "yes", "no", and "unrelated", respectively. The worker is then required to answer a prerequisite question correctly to proceed to the formal survey. The formal survey is composed of ten questions, including two hidden qualification checking questions. Before publishing on the AMT, two Ph.D. students

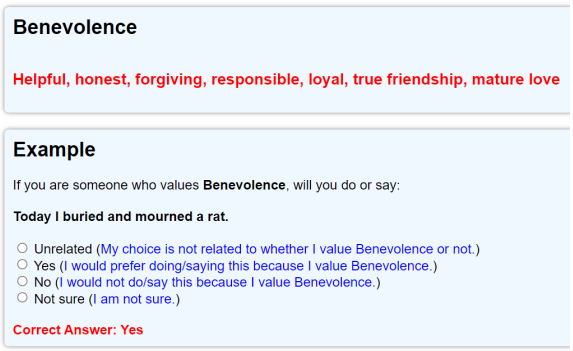

Figure 3: Value-aspect attitude annotation in AMT.

prepared the qualification questions by annotating a small subset of the curated scenarios. Their agreed samples (100 in total) were randomly inserted into the survey for worker selection. The selection procedure was done in the value dimensions with more scenarios to get a large pool of qualified workers and a relatively balanced final dataset across different values. The complete Mechanical Turk interface is attached in the Appendix for reference.

A total of 681 experienced AMT workers participated in our VALUENET annotation. 443 of them passed the qualification test. Each scenario is assigned to four different workers. The original inter-annotator agreement is 64.9%, and the Fleiss' kappa score (Fleiss 1971) among the workers is 0.48, which considers the possibility of the agreement by chance. Keeping the scope of VALUENET in commonly-agreed attitudes towards social scenarios, we only retain the samples with three or more agreements. Figure 4 shows the sample size of each value split and their label distribution.

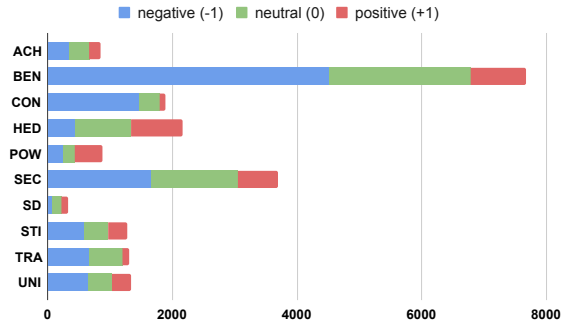

Figure 4: The sample number and label distribution of each value split in the VALUENET.

The data is split into the train (75%), valid (15%), and test (10%). Similar to the polarity in sentiment analysis (Kouloumpis, Wilson, and Moore 2011), we quantify the annotated labels into numerical values: yes (positive): +1, no (negative): -1, unrelated (neutral): 0. We denote the numerical values as **utility** to describe the effect of a scenario on one's value. In other words, for people who appreciate a certain value, actions with a higher utility in this value dimension would be more desirable to them.

VALUENET	train	valid	test	total
# samples	16,030	3,206	2,138	21,374
average # tokens	12.05	12.09	12.26	12.07
unique # tokens	12,452	5,292	4,112	14,143

Table 1: Statistics of the VALUENET dataset.

Table 1 shows more statistical details about the VALUENET dataset. In total, we collected 21,374 samples covering a wide range of scenarios in daily social life.

Value Modeling

We experiment using Transformer-based pre-trained language models for modeling human values from the VALUENET dataset.

Task Formalization

Given a social scenario s, we wish to learn a value function that models the utility distribution of s from the ten Schwartz value dimensions: $\mathbf{V}(s) = [V_{\text{SEC}}(s), V_{\text{POW}}(s), V_{\text{ACH}}(s), V_{\text{HED}}(s), V_{\text{STI}}(s), V_{\text{SD}}(s), V_{\text{UNI}}(s), V_{\text{BEN}}(s), V_{\text{CON}}(s), V_{\text{TRA}}(s)]$, where $V_{\$\text{VALUE}}(\cdot) \in [-1, 1]$ and $V_{\$\text{VALUE}}(\cdot) \in \mathbb{R}$.

Model

Pre-trained language model variants: BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), DistilBERT (Sanh et al. 2019), BART (Lewis et al. 2019) are investigated for learning the value function. A custom input format constructed as '`[CLS] [$VALUE] s`' is fed into a Transformer encoder, *i.e.,*

$$V_{\$\text{VALUE}}(\text{s}) = \text{TRM}(\texttt{[CLS] [\$VALUE] s}), \qquad (1)$$

where TRM denotes the Transformer encoder, `[CLS]` is the special token for regression or classification, and `[$VALUE]` are special tokens we define to prompt the language models the value dimension we are interested in (Li and Liang 2021; Brown et al. 2020; Le Scao and Rush 2021). In order to get the ten-dimensional output $\mathbf{V}(s)$, a batch size of 10 is forwarded through the model. For the BERT, DistilBERT, and RoBERTa, a regression head is put on top of the models and they are trained with the Mean Squared Error (MSE) loss. We use the regression model with $sigmoid$ activation to get a continuous estimation of the utility in the range of $[-1, 1]$. To evaluate the effect of different loss functions, we train the BART model with three output classes and the cross-entropy loss.

Result and Analysis

The learning performance of using fastText[4] (Joulin et al. 2017) and Transformer variants are reported in Table 2. All Transformers are trained for 40 epochs with a learning rate

[4]https://github.com/facebookresearch/fastText

		F1(-1)	F1(0)	F1(1)	P(-1)	P(0)	P(1)	R(-1)	R(0)	R(1)	Acc.↑	MSE↓
VALUENET (original)	fastText	0.70	0.46	0.43	0.65	0.47	0.55	**0.76**	0.44	0.35	0.58	0.66
	BERT	**0.73**	0.50	0.51	0.72	0.46	0.71	0.74	0.55	0.39	0.61	0.39
	DistilBERT	0.71	0.52	0.47	0.74	0.45	0.69	0.68	0.62	0.36	0.60	**0.37**
	RoBERTa	0.65	0.51	0.34	0.74	0.40	0.71	0.58	0.69	0.22	0.55	0.41
	BART	0.00	**0.76**	0.54	0.00	0.70	0.60	0.00	**0.83**	**0.49**	**0.67**	0.52
VALUENET (balanced)	fastText	0.70	0.48	0.43	0.64	0.50	0.54	**0.76**	0.45	0.36	0.59	0.68
	BERT	0.67	0.48	0.51	0.73	0.42	0.61	0.62	0.58	0.43	0.57	0.40
	DistilBERT	0.66	0.49	0.50	0.74	0.41	0.61	0.60	0.60	0.43	0.57	0.40
	RoBERTa	0.65	0.51	0.34	0.74	0.40	0.71	0.58	0.69	0.22	0.55	0.41
	BART	0.00	0.75	0.51	0.00	0.72	0.57	0.00	0.77	0.47	0.65	0.55
VALUENET (augmented)	fastText	0.58	0.52	0.29	0.72	0.40	0.65	0.49	0.75	0.18	0.52	0.59
	BERT	0.67	0.55	0.41	0.78	0.43	**0.78**	0.58	0.76	0.28	0.58	0.38
	DistilBERT	0.68	0.57	0.41	**0.79**	0.44	**0.78**	0.59	0.78	0.28	0.60	0.38
	RoBERTa	0.70	0.56	0.41	0.78	0.45	0.75	0.64	0.74	0.28	0.61	0.40
	BART	0.00	0.74	**0.57**	0.00	**0.75**	0.49	0.00	0.73	0.66	0.64	0.46

Table 2: Value modeling performance in the VALUENET dataset. **Bold** items are the best in each metric column.

Acc.	ACH	BEN	CON	HED	POW	SEC	SD	STI	TRA	UNI
VALUENET (original)	**0.56**	**0.68**	0.82	**0.63**	0.35	**0.52**	0.45	**0.58**	0.60	**0.51**
VALUENET (balanced)	0.53	0.58	**0.83**	**0.63**	**0.41**	0.50	0.42	0.53	0.61	0.50
VALUENET (augmented)	0.48	0.66	0.82	0.58	0.33	0.47	**0.48**	0.49	**0.64**	0.42

Table 3: Accuracies of the BERT (Devlin et al. 2018) value model across different value dimensions in the VALUENET dataset.

of 5e−6. The prediction precision, recall, **F**1 score, and accuracy for regression models are computed by the utility rounded to the nearest integer.

In general, pre-trained language models perform better than the fastText baseline. However, there is not a noticeable difference between the Transformer variants. The prediction accuracy of BART is the highest among all models because it is explicitly trained for classification purposes. BERT and DistilBERT get the lowest MSE in terms of regression performance.

Observing the sample imbalance across different value splits and labels (Figure 4), we release another two versions of VALUENET: VALUENET (balanced) and VALUENET (augmented). The original dataset is balanced by subsampling the negative and neutral data of the largest value split (BENEVOLENCE). Moreover, we augment the neutral class of the original VALUENET by assigning AMT results with less worker agreement to "unrelated". Data distribution of the balanced and augmented versions of VALUENET are illustrated in the Appendix. By analyzing the prediction accuracy in different value splits (Table 3), we find that reducing the sample number of BENEVOLENCE hurts the model performance in that dimension. Looking at the **F**1 score of each class in Table 2, we conclude that augmenting the neutral class improves the **F**1(0) but reduces **F**1(1) and **F**1(-1). We leave it a future work to further improve the value modeling performance.

In the next sections, we show how the learned value function could benefit EQ-related tasks and help build a value-driven dialogue system.

PERSONA-CHAT

As values are closely related to one's personality, we first assess our value model on a personalized dialogue dataset: PERSONA-CHAT (Zhang et al. 2018). The PERSONA-CHAT dataset contains multi-turn dialogues conditioned on personas. Each persona is encoded by at least 5 sentences of textual description, termed a profile. Example profile sentences are "I like to ski", "I enjoying walking for exercise", "I have four children", *etc.* The dataset is composed of 8,939 dialogues for training, 1,000 for validation, and 968 for testing. It also provides *revised* personas by rephrasing, generalizing or specializing the *original* ones. The dataset we use for experiments is public available in ParlAI[5].

Task Formalization

Given the agent's self persona profile $\mathbf{p} = [p_1, p_2, ..., p_N]$ and the dialogue history up to the t-th turn $\mathbf{h}_t^s = (x_1^u, x_1^s, ..., x_t^u)$, x_i^u is the i-th utterance by Person 1 played by the user, x_i^s is the i-th utterance by Person 2 played by the system, we evaluate the model's performance on predicting the next utterance x_t^s.

Model

A decoder-only Transformer-based model is used to estimate the generation distribution $p_\theta(x_t^s \mid \mathbf{h}_t^s, \mathbf{p})$, where θ is the model parameter. Following the practice proposed in Guo et al. (2018), the model is firstly trained with Maximum Likelihood Estimation (MLE) to ensure generating fluent responses. Then we took an interleaving of supervised

[5]https://parl.ai/projects/convai2/

Model	Original			Revised		
	Hits@1(%) ↑	**Ppl.**↓	**F1(%)** ↑	**Hits@1(%)** ↑	**Ppl.**↓	**F1(%)** ↑
SEQ2SEQ-ATTN	12.5	35.07	16.82	9.8	39.54	15.52
$\mathcal{P}^2$BOT (Liu et al. 2020)	–	15.12	19.77	–	18.89	19.08
GPT2 (MLE) (Radford et al. 2019)	$14.51_{[0.05]}$	$17.23_{[0.03]}$	$18.74_{[0.01]}$	$10.31_{[0.07]}$	$20.64_{[0.11]}$	$18.29_{[0.05]}$
GPT2 + Value (Ours)	$16.44_{[0.10]}$	$16.83_{[0.06]}$	$18.76_{[0.02]}$	$12.19_{[0.03]}$	$19.98_{[0.06]}$	$17.88_{[0.05]}$
DialoGPT (MLE) (Zhang et al. 2019)	$20.20_{[0.04]}$	$14.38_{[0.05]}$	$20.16_{[0.04]}$	$15.80_{[0.03]}$	$17.35_{[0.05]}$	$19.08_{[0.08]}$
DialoGPT + Value (Ours)	$\mathbf{20.97}_{[0.08]}$	$\mathbf{13.84}_{[0.03]}$	$\mathbf{20.22}_{[0.01]}$	$\mathbf{18.83}_{[0.03]}$	$\mathbf{17.01}_{[0.03]}$	$\mathbf{19.79}_{[0.10]}$

Table 4: Next Utterance Prediction Performance on PERSONA-CHAT (Zhang et al. 2018). We report the standard deviation $[\sigma]$ (across 5 runs) of the models we trained.

training (MLE) and reinforcement learning. We use the REINFORCE policy gradient algorithm (Williams 1992) in our experiment, and the reward assignment is described as following.

Denote $\mathbf{V}(p_i)$ and $\mathbf{V}(\hat{x}_i^s)$ to describe the estimation of the agent's value from its profile sentence p_i and generated response $\hat{x}_i^s$, respectively. We want the reward to promote the alignment of the agent's profile and utterances in the value space. For instance, if the agent has profile '*I like venture*' and '*I have a dog*', and it says '*I plan to ski this weekend*' and also '*Do you like skiing*'. Both utterances should be aligned with the first persona. Here we propose a simple yet effective searching algorithm (Algorithm 1) to find a match between $[\mathbf{V}(p_1), \mathbf{V}(p_2), ..., \mathbf{V}(p_N)]$ and $[\mathbf{V}(\hat{x}_1^s), \mathbf{V}(\hat{x}_2^s), ..., \mathbf{V}(\hat{x}_T^s)]$ and return a reward R. N is the number of profile sentences and T is the length of the generated dialogue. $\mathbf{V}$ is normalized to ensure $|r_t| \leq 1$. Intuitively, the discount argument γ prevents the language model from repeating the same fact in the agent's profile.

Setup

We evaluate the same generative model in both generation and ranking settings. In the response ranking setup, the candidates are scored with their log-likelihood. For the GPT-2 (Radford et al. 2019) and DialoGPT (Zhang et al. 2019) we have finetuned, we train them for 5k steps with a training batch size of 8. The learning rate is set to 2e−6. For an illustration of computational requirements, the training with MLE on 4 NVIDIA Tesla V100 takes ∼1 hours, and the reinforcement learning takes ∼30 minutes.

Result and Analysis

Following Zhang et al. (2018) and Liu et al. (2020), we report the **Hits**@1, **Perplexity** and **F**1 to evaluate the methods in Table 4. By the submission of this paper, $\mathcal{P}^2$BOT (Liu et al. 2020) is the state-of-the-art model reported in this task. We also include a generative baseline using SEQ2SEQ with attention mechanism (Bahdanau, Cho, and Bengio 2014) for comparison. As observed, in terms of all the metrics we evaluated, finetuning GPT2 or DialoGPT2 models with our value function provides a significant performance boost compared to simply training them with MLE. Our DialoGPT + Value model achieves new state-of-the-art performance on perplexity and **F**1.

Algorithm 1: Personalized Dialogue Value Matching

Input: $[\mathbf{V}(p_1), ..., \mathbf{V}(p_N)], [\mathbf{V}(\hat{x}_1^s), ..., \mathbf{V}(\hat{x}_T^s)]$
Output: reward R

for $t = 1, 2, ..., T$ **do**
  $r_t \leftarrow -1$
  $m_t \leftarrow -1$
  for $i = 1, 2, ..., N$ **do**
    if $\mathbf{V}(p_i) \cdot \mathbf{V}(\hat{x}_t^s) > r_t$ **then**
      $r_t \leftarrow \mathbf{V}(p_i) \cdot \mathbf{V}(\hat{x}_t^s)$
      $m_t \leftarrow i$
    end if
  end for
end for
$\gamma_i \leftarrow 1, i = 1, 2, ..., N$
for $t = 1, 2, ..., T$ **do**
  $\gamma_{m_t} \leftarrow \gamma_{m_t} + 1$
end for
$R \leftarrow 0$
for $t = 1, 2, ..., T$ **do**
  $R \leftarrow R + \text{sign}(r_t) \cdot |r_t|^{\text{sign}(r_t) \cdot \gamma_{m_t}}$
end for
return R/N

EMPATHETICDIALOGUES

EMPATHETICDIALOGUES (Rashkin et al. 2019) provides 25k conversations grounded in emotional situations. It aims to test the dialogue system's capability to produce empathetic responses. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion, with a listener responding. In this section, we demonstrate how we could leverage VALUENET to improve the emotion classification accuracy and further improve the empathetic response generation.

Emotion Classification

An auxiliary task that is highly related to empathetic dialogue generation is emotion classification. In EMPATHETICDIALOGUES, each situation is written in association with a given emotion label. A total of 32 emotion labels were annotated to cover a broad range of positive and negative emotions.

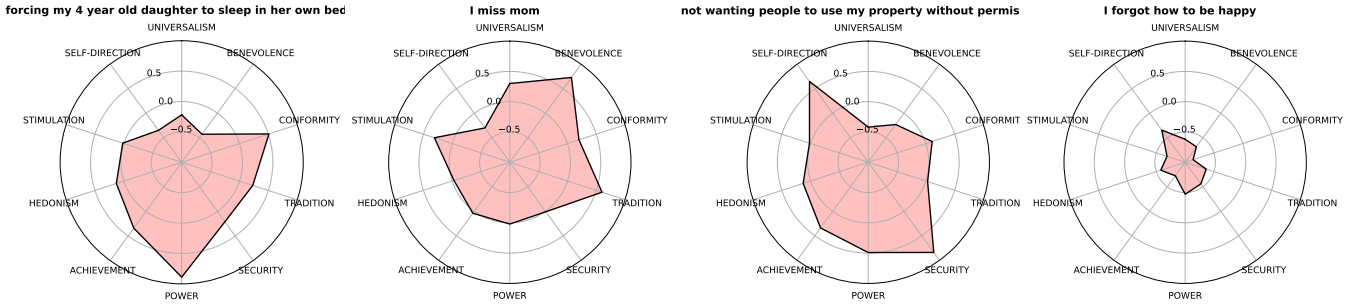

Figure 5: Value visualization of example utterances/scenarios.

Model Given the situation context s, a pre-trained BERT model encodes s and gets the sentence representation from its pooling layer of the [CLS] token. The same context is parsed by our pre-trained value model to get a ten-dimensional vector, which serves as an additional feature for the classification:

$$\begin{aligned} h_s &= \text{BERT}(s), \\ v_s &= \mathbf{V}(s), \\ e &= softmax(\mathbf{W} \cdot ([h_s; v_s]) + \mathbf{b}), \end{aligned} \tag{2}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters.

Result We compare the performance between our implementation and the baseline that directly applies the BERT model for emotion classification. As shown in Table 5, the additional value information benefits emotion classification from both the DistilBERT and BERT models. Our method obtains a **relative** improvement of 5.2% on DistilBERT and 6.4% on BERT.

Model	**Accuracy** (σ)	
fastText	$42.27 \pm 0.3\%$	
DistilBERT	$41.81 \pm 0.2\%$	
DistilBERT + Value	$43.98 \pm 0.2\%$	$+\mathbf{2.17}\%$
BERT	$42.93 \pm 0.1\%$	
BERT + Value	$\mathbf{45.67} \pm 0.3\%$	$+\mathbf{2.74}\%$

Table 5: Emotion classification performance in EMPATHETICDIALOGUES (Rashkin et al. 2019).

Empathetic Dialogue Generation

We further check whether our value model helps the empathetic dialogue generation. EMPATHETICDIALOGUES applies PREPEND-K, a strategy to add supervised information to data, when predicting the utterance given the dialogue history and the situation. We apply the strategy of prepending the top-k emotion labels for dialogue generation. The top predicted label from the classifiers of emotion is prepended to the beginning of the token sequence as encoder input, as below:

- **Original**: "I finally got promoted!"
- **Prepend-1 emotion**: "*proud* I finally got promoted!"

Result The results are shown in Table 6. As observed, prepending emotion tokens provides extra context and improves the generation performance of GPT2 and DialoGPT. Since incorporating value improves the emotion classification accuracy, it further improves the generation quality.

Model	**Ppl.**$\downarrow$
EmoPrepend-1 (Rashkin et al. 2019)	24.30
GPT	14.74
GPT + Emotion (w/o Value)	14.46
GPT + Emotion (w/ Value)	14.01
DialoGPT	13.48
DialoGPT + Emotion (w/o Value)	12.32
DialoGPT + Emotion (w/ Valued)	**12.12**

Table 6: Empathetic dialogue generation in EMPATHETIC-DIALOGUES Rashkin et al. (2019). EmoPrepend-1: input prepending emotion from an external classifier.

Value Profiling

For a more comprehensive understanding, we visualize the 10-dimensional value of four example scenarios in Figure 5. As shown, the value model provides a numerical speaker profile. For instance, saying "forcing my daughter to sleep in her own bed" implies that the speaker values power and conformity; saying "I miss mom" implies that the speaker values benevolence; saying "not wanting people to use my property without permissions" implies the speaker is self-directed and values security. The last example "I forgot how to be happy" results a small radar graph. It suggests that even the model could predict the overall polarity pretty well, there is still space to improve its capability of distinguishing different values.

Conclusion

We introduce a new dataset for human value modeling, VALUENET, which contains 21,374 scenarios in ten distinct human values. We also apply the learned value model from VALUENET to several EQ-related dialogue tasks. Our experiments show our approach and dataset provide a new way to control the dialogue system speaking style and numerically estimate one's value preference. We hope that our results and dataset will stimulate more research in the important direction of building human-value-driven dialogue systems.

Ethical Statement

The original purpose of introducing Schwartz values is to identify individual values that are recognized across cultures, which is based on surveys conducted among 82 countries. This motivates us to seek commonly agreed attitudes on social scenarios. Considering model reasoning capability and scalability, we follow the practice in recent commonsense works and provide one-sentence text descriptions to annotators. However, we acknowledge the limitation of this approach lacking external contexts such as culture and language diversity. While some scenarios might have higher levels of agreement across cultures, others might have dramatic variations. As a starting point, our study focuses on the value modeling of English-speaking cultures represented within North America. Extending this formalism to other countries and non-English speaking cultures remains a compelling area of future research. Bearing this in mind, we hope this paper could stimulate research on building scalable value-aligned AI.

References

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

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*.

Colby, K. M.; Weber, S.; and Hilf, F. D. 1971. Artificial paranoia. *Artificial Intelligence*, 2(1): 1–25.

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

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, 187–208. Springer.

Fedorenko, D.; Smetanin, N.; and Rodichev, A. 2018. Avoiding echo-responses in a retrieval-based conversation system. In *Conference on Artificial Intelligence and Natural Language*, 91–97. Springer.

Fishburn, P. C. 1970. Utility theory for decision making. Technical report, Research analysis corp McLean VA.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670. Online: Association for Computational Linguistics.

Gabriel, S.; Bhagavatula, C.; Shwartz, V.; Bras, R. L.; Forbes, M.; and Choi, Y. 2021. Paragraph-level commonsense transformers with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2470–2481. Online: Association for Computational Linguistics.

Guo, J.; Lu, S.; Cai, H.; Zhang, W.; Yu, Y.; and Wang, J. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

He, H.; Balakrishnan, A.; Eric, M.; and Liang, P. 2017. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1766–1776. Vancouver, Canada: Association for Computational Linguistics.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Hsu, C.-C.; Chen, S.-Y.; Kuo, C.-C.; Huang, T.-H.; and Ku, L.-W. 2018. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Association for Computational Linguistics.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*.

Le Scao, T.; and Rush, A. M. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2627–2636.

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, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Liang, Y.; He, L.; and Anthony'Chen, X. 2021. Human-Centered AI for Medical Imaging. *Artificial Intelligence for Human Computer Interaction: A Modern Approach*, 539–570.

Liu, Q.; Chen, Y.; Chen, B.; Lou, J.-G.; Chen, Z.; Zhou, B.; and Zhang, D. 2020. You Impress Me: Dialogue Generation via Mutual Persona Perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1417–1427. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lourie, N.; Bras, R. L.; and Choi, Y. 2021. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

McKeown, G.; Valstar, M. F.; Cowie, R.; and Pantic, M. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*, 1079–1084. IEEE.

Okuda, T.; and Shoda, S. 2018. AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal*, 54(2): 4–8.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536. Florence, Italy: Association for Computational Linguistics.

Qiu, L.; Liang, Y.; Zhao, Y.; Lu, P.; Peng, B.; Yu, Z.; Wu, Y. N.; and Zhu, S.-C. 2021a. SocAoG: Incremental Graph Parsing for Social Relation Inference in Dialogues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 658–670. Online: Association for Computational Linguistics.

Qiu, L.; Zhao, Y.; Liang, Y.; Lu, P.; Shi, W.; Yu, Z.; and Zhu, S.-C. 2021b. Towards Socially Intelligent Agents with Mental State Transition and Human Utility. *arXiv preprint arXiv:2103.07011*.

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

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381. Florence, Italy: Association for Computational Linguistics.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019. Social IQa: Commonsense Reasoning about Social Interactions. 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)*, 4463–4473. Hong Kong, China: Association for Computational Linguistics.

Schwartz, S. H. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1): 2307–0919.

Schwartz, S. H.; Cieciuch, J.; Vecchione, M.; Davidov, E.; Fischer, R.; Beierlein, C.; Ramos, A.; Verkasalo, M.; Lönnqvist, J.-E.; Demirutku, K.; et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4): 663.

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1): 36–45.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3): 229–256.

Yu, D.; Sun, K.; Cardie, C.; and Yu, D. 2020. Dialogue-Based Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4927–4940. Online: Association for Computational Linguistics.

Yuan, W.; and Li, Z. 2017. Development of a human-friendly robot for socially aware human-robot interaction. In *2017 2nd International Conference on Advanced Robotics and Mechatronics (ICARM)*, 76–81. IEEE.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.

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

Zhao, R.; Romero, O. J.; and Rudnicky, A. 2018. SOGO: a social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 239–246.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, 46(1): 53–93.