Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems

Weiwei Sun[1*] Shuo Zhang[2*] Krisztian Balog[3] Zhaochun Ren[1†]
Pengjie Ren[1] Zhumin Chen[1] Maarten de Rijke[4,5]
[1]Shandong University, Qingdao, China
[2]Bloomberg, London, United Kingdom [3]University of Stavanger, Stavanger, Norway
[4]University of Amsterdam [5]Ahold Delhaize Research, Amsterdam, The Netherlands
sunnweiwei@gmail.com,szhang611@bloomberg.net,krisztian.balog@uis.no
{zhaochun.ren,chenzhumin}@sdu.edu.cn,jay.ren@outlook.com,m.derijke@uva.nl

ABSTRACT

Evaluation is crucial in the development process of task-oriented dialogue systems. As an evaluation method, user simulation allows us to tackle issues such as scalability and cost-efficiency, making it a viable choice for large-scale automatic evaluation. To help build a human-like user simulator that can measure the quality of a dialogue, we propose the following task: simulating user satisfaction for the evaluation of task-oriented dialogue systems. The purpose of the task is to increase the evaluation power of user simulations and to make the simulation more human-like. To overcome a lack of annotated data, we propose a user satisfaction annotation dataset, User Satisfaction Simulation (USS), that includes 6,800 dialogues sampled from multiple domains, spanning real-world e-commerce dialogues, task-oriented dialogues constructed through Wizard-of-Oz experiments, and movie recommendation dialogues. All user utterances in those dialogues, as well as the dialogues themselves, have been labeled based on a 5-level satisfaction scale. We also share three baseline methods for user satisfaction prediction and action prediction tasks. Experiments conducted on the USS dataset suggest that distributed representations outperform feature-based methods. A model based on hierarchical GRUs achieves the best performance in in-domain user satisfaction prediction, while a BERT-based model has better cross-domain generalization ability.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Human-centered computing → Human computer interaction (HCI)**.

[*]Equal contribution.
[†]Corresponding author.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463241

KEYWORDS

User simulation, task-oriented dialogue, conversational recommendation, conversational information access

ACM Reference Format:
Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3404835.3463241

1 INTRODUCTION

Task-oriented systems are developed to help users solve a specific task as efficiently as possible [56]. Evaluation is a crucial part of the development process of task-oriented dialogue systems. For evaluating the performance of each module of a dialogue system, human evaluation, user satisfaction modeling, corpus-based approaches, and user simulation have all been leveraged [13]. Human evaluation through in-field experiments [6, 27] or crowd-sourcing [25] is considered to reflect the overall performance of the system in a real-world scenario, but it is intrusive, time-intensive, and does not scale [13]. User satisfaction modeling can be an alternative; it aims to automatically estimate user satisfaction based on human-machine interaction log data, but still requires human involvement. To evaluate a dialogue system fully automatically, offline evaluation based on test sets is commonly used. However, this method is limited to a single turn and does not inform us about the overall usefulness of the system or about users' satisfaction with the flow of the dialogue [57]. Therefore, evaluation results of offline methods have limited consistency with the results of human evaluation. Simulation-based evaluation methods address the issues listed above; they are a viable choice for large-scale automatic evaluation [13]. User simulations can be used to evaluate functionalities of dialogue systems and they can serve as an environment to train reinforcement learning-based systems [13], leveraging agenda-based [42] or model-based simulation [3]. Building human-like user simulation is still an open challenge [22].

To bridge the gap between human evaluation and user simulation, we attempt to combine user simulation with user satisfaction (cf. Figure 1). To this end, we first look into existing task-oriented dialogues and carry out a user study to investigate the characteristics of user satisfaction. We arrive at two main observations:

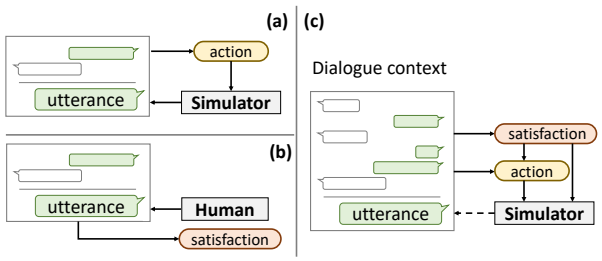

Figure 1: (a) Previous work on user simulation; (b) previous work on user satisfaction prediction; (c) our proposed task: simulating user satisfaction for evaluating task-oriented dialogues systems. We leave utterance generation (dotted line) as future work.

(1) *User dissatisfaction is mainly caused by the system's failure in meeting the user's needs.* Specifically, 36% of the conversations are labeled as *very dissatisfied* because the system does not understand the user's needs, and 43% are because the system understands the user's problems but cannot provide proper solutions. Figure 2 illustrates the scenario. (2) *Different degrees of satisfaction result in different sequences of user actions.* For example, the right-side user in Figure 2 may switch to customer service or explain further when encountering the same failed system reply in the context of different emotions. We convert this intuition to a hypothesis that we verify by checking the records in the corpus. When faced with a dialogue system's failure in understanding user needs, about 17.1% of all users will switch to manual customer service, and about 64.3% and 9.7% will continue by providing additional information, or quit the conversation, respectively. This observation suggests that user simulation should work differently in different user satisfaction scenarios.

Informed by the observations just listed, we propose a novel task: *to simulate user satisfaction for the evaluation of task-oriented dialogue systems.* Figure 1 illustrates the main difference between our task and previous work. We extend the evaluation capability of user simulations and make the simulation more human-like by incorporating user satisfaction prediction and user action prediction.

To facilitate research on user satisfaction simulation, we develop a user satisfaction annotation dataset, *User Satisfaction Simulation* (USS). We invite 40 annotators to label both the dialogue level and exchange level user satisfaction of 5 commonly used task-oriented dialogue datasets in different domains. This results in a dataset of 6,800 dialogues, where each individual user utterance, as well as each complete dialogue, is labeled on a 5-point satisfaction scale. Each dialogue is labeled by 3 annotators; the expert ratings are highly correlated, with a Fleiss Kappa score of 0.574. The User Satisfaction Simulation (USS) dataset shares some characteristics with existing datasets for user satisfaction, but also differs in important ways (see Table 1): (1) Our user satisfaction labeling occurs before the user utterance, and is based on the dialogue context between user and system instead of the satisfaction expressed in the user's utterance. (2) The USS dataset includes multiple domains, such as e-commerce, reservations, recommendations, etc. (3) The USS dataset exceeds existing user satisfaction data in scale.

We share three baseline approaches to perform satisfaction prediction and user action prediction based on the newly collected

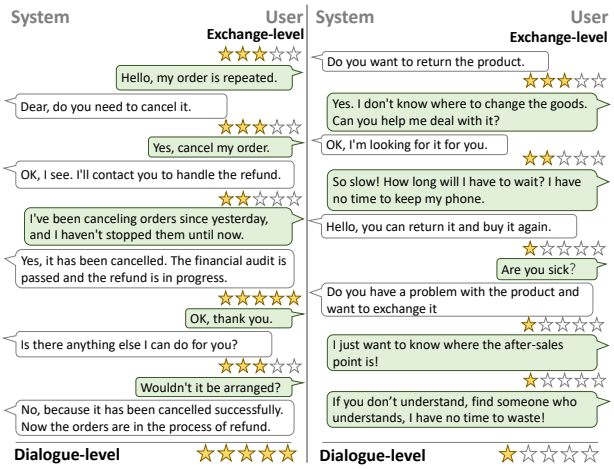

Figure 2: Two examples of dialogues in the JDDC dataset [11], with different degrees of user satisfaction. The right-side system fails to understand the user's needs, and causes the user to be dissatisfied emotions and have a poor user experience. The left-side dialogue demonstrates an opposite case.

data in USS: a feature-based method, a hierarchical GRU-based method, and a BERT-based method. Experimental results suggest that distributed representations outperform feature-based methods. The hierarchical GRU-based method achieves the best performance in in-domain user satisfaction prediction, while the BERT-based method has a better cross-domain generalization ability thanks to the pre-training. We also show that the BERT-based method achieves state-of-the-art performance on the action prediction task.

In summary, this paper makes the following contributions: (1) We propose the novel task of simulating user satisfaction for the evaluation of task-oriented dialogue systems. (2) We collect and share a dataset, USS, that includes 6,800 annotated dialogues in multiple domains. (3) We introduce three baseline methods for the tasks of satisfaction prediction and action prediction using the USS dataset.

2 RELATED WORK

Unlike chitchat systems, which focus on conversing with human on open domains, task-oriented dialogue systems aim to complete specific tasks for user [29, 51]. Task-oriented dialogue systems can be divided into module-based and end-to-end-based methods [22]. The former decomposes the dialogue system into four stages: language understanding, dialogue state tracking, dialogue policy learning, and response generation. Recently, each stage in the module-based task-oriented dialogue systems has received increased attention [20, 37, 38, 51, 53, 55]. End-to-end task-oriented dialogue systems rely on neural dialogue generation, which has received a lot of attention in recent years [2, 4, 56]. Among all these approaches, sequence-to-sequence structure neural generation models [10, 24, 28, 31, 46, 49] have been proved to be capable in multiple dialogue systems with promising performance.

Evaluation is a crucial part of the development process of task-oriented dialogue systems. Corpus-based approaches, user simulation, and user satisfaction modeling have all been leveraged [57]

Table 1: Available datasets related to our task. AU/BU is short for After Utterance/Before Utterance.

Dataset	Year	Domain	#Dialog	#Turns	Type	Level
LEGO [45]	2012	Bus	347	9,083	AU	5
IARD [9]	2020	Movie	336	2,203	AU	2
Alexa [8]	2020	Booking	3,129	20,167	AU	5
MHCH [34]	2020	E-commerce	7,500	75,548	BU	2
USS (Ours)	2021	Multiple	6,800	99,569	BU	5

for evaluating the performance of a task-oriented dialogue system. Offline evaluation based on test sets is commonly used, but it is limited in scope to a single turn and does not inform us about the overall usefulness of the system or about users' satisfaction with the flow of the dialogue [57]. Employing simulation-based evaluation can tackle the above issues and become one viable choice for large-scale automatic evaluation [13]. User simulators are tools that are designed to simulate the user's behavior, which can be used to train the dialogue manager in an offline environment [13] or to evaluate the dialogue policy [42]. Eckert et al. [15] propose the first statistical user simulator. Scheffler and Young [43] propose a graph-based model. Georgila et al. [18] use a Markov Model, and a hidden Markov model has been proposed by Cuayáhuitl et al. [12]. Schatzmann et al. [42] propose an agenda-based user simulator, which represents the user state elegantly as a stack of necessary user actions, called the agenda. Zhang and Balog [57] evaluate conversational recommender systems via an agenda-based user simulator. Recent work employs neural approaches, esp. sequence-to-sequence models [3, 26]. As far as we know, no previous study explicitly models the user satisfaction in user simulations. Unlike previous work, we are the first to incorporate user satisfaction into user simulation to make the simulation more human-like.

Next to user simulations, user satisfaction modeling is the other evaluation method that is based on the idea that the usability of a system can be approximated by the satisfaction of its users [13]. Ultes et al. [48] note the impracticability of having a user rate a live dialogue. Thus, automatic prediction can be an alternative. Walker et al. [50] propose the PARADISE framework, which estimates user ratings on the dialogue level. Evaluation methods that estimate user satisfaction at the exchange level have also been proposed [16, 19, 21]. They yield more fine-grained predictions and are especially useful for online dialogue breakdown detection. Schmitt and Ultes [44] propose Interaction Quality (IQ) to assign user ratings by experts instead of real users. Bodigutla et al. [7] introduce the Response Quality (RQ) scheme to improve generalizability to multiple-domain conversations.

Unlike previous work on user satisfaction modeling, we simulate the user satisfaction changes without human involvement.

3 TASK FORMULATION

To formulate the task of simulating user satisfaction, we first carry out a user study to explore the characteristics of user satisfaction in task-oriented dialogues. Specifically, we invite 12 experts and let each expert annotate 20 dialogues sampled from the JDDC dataset; we used the JDDC dataset since it is more realistic than data constructed by the Wizard-of-Oz approach. We ask each expert to score the user satisfaction for each dialogue turn and the entire conversation. In addition, a rational explanation is requested. We ask the experts to judge the user action changes after a change in satisfaction. Based on this study, we answer the following questions:

(1) *What causes the user's dissatisfaction?* We collect the results and find that, although annotators are satisfied with the system overall, about 12% of the dialogue turns are labeled as unsatisfying. This indicates that there are fluctuations in user satisfaction when interacting with the system. We analyze the annotators' explanations and find that the main reason for dissatisfaction relates to the system's failure to understand the user's needs or handling the user's requests. Specifically, 36% of all conversations labeled as *very dissatisfied* are because *the system does not understand the user's needs*, whereas 43% are because *the user does not approve the system's response*. In 64% of the data, users had a bad user experience because *the system was not professional enough or did not respond in time*. Figure 2 illustrates the scenario where the system does not understand the user's needs and causes low user satisfaction. (2) *How does user satisfaction influence the user's behavior?* Different degrees of satisfaction result in different sequences of user actions. Specifically, when encountering a failure in the a dialogue system's understanding of user needs, about 17.1% of all users *switch to manual customer service*, and about 64.3% and 9.7% continue by *providing additional information*, or *quit the conversation*, respectively. Figure 2 shows an example, where the right-side user switches to customer service or explains further when encountering the same failed system reply in light of different degrees of satisfaction. Apart from user actions, we also observe changes such as attitude and information-seeking goal.

The above observations indicate that predicting the fluctuations of user satisfaction during interaction is non-trivial. Thus, we formulate our research task, i.e., to *simulate user satisfaction for the evaluation of task-oriented dialogue systems*.

This simulation task focuses on the prediction of the next user action as well as user satisfaction. Suppose that we have a dataset $\mathcal{D} = \{(U_i, a_i, s_i)\}_{i=1}^{N}$, where for all $i \in [1, N]$, U_i is the dialogue context, a_i is the next-turn user action, and s_i denotes user satisfaction. The task objective is to learn a classification model $P(a, s \mid U)$ from $\mathcal{D}$, and thus given a dialogue context U, it predicts the next-turn user action a and user satisfaction s based on $P(a, s \mid U)$. The purpose of the task to increase the evaluation power of user simulations and to make the simulation more human-like by incorporating the user's potential changes in satisfaction in a simulator.

4 CONSTRUCTING A TEST COLLECTION

We propose a user satisfaction annotation dataset, *User Satisfaction Simulation* (USS). Below, we detail the creation of the dataset. We divide this section into 3 phases: data preparation, user satisfaction assessment, and measures and disclaimers.

4.1 Data preparation

The USS dataset is based on five benchmark task-oriented dialogue datasets: JDDC [11], Schema Guided Dialogue (SGD) [41], MultiWOZ 2.1 [17], Recommendation Dialogues (ReDial) [33], and Coached Conversational Preference Elicitation (CCPE) [40]. We first distinguish the user's emotion in the conversation by a classifier trained on annotated reddit data (weibo for Chinese), and then

filter out all conversations that do not show negative emotions (i.e., anger, disgust, fear, sadness).

(1) JDDC is a large-scale, real-world Chinese e-commerce conversation corpus with over 1 million multi-turn dialogues. We first classify the conversation into 11 types according to the type of transaction, e.g., delivery, return, invoice, etc. Then, we sample 300 dialogue sessions from each type, for a total of 3,300 conversations. The JDDC data set provides the action of each user utterance, including 234 categories. We compress them into 12 categories based on a manually defined classification method. (2) SGD is a dataset consisting of over 20K annotated task-oriented conversations between a human and a virtual assistant spanning 16 domains. MultiWOZ 2.1 is a multi-domain dialogue dataset spanning 7 distinct domains and containing over 10K dialogues. We sample 1,000 conversations from the two datasets. We directly use the action annotation from the original datasets. The SGD has 12 actions, and MultiWOZ has 21 actions. (3) ReDial is an annotated dataset consisting of over 10K conversations, where users recommend movies to each other. We sample 1,000 dialogues. Since the original dataset does not provide actions, we use the action annotation provided by IARD [9]. (4) CCPE is a dataset consisting of 502 dialogues with 12K annotated utterances between a user and an assistant discussing movie preferences. We sample 300 dialogues from the CCPE dataset and used the actions provided by the original dataset.

4.2 User satisfaction assessment

We hired 40 annotators to annotate exchange-level and dialogue-level user satisfaction levels of each conversation with five levels (1–5). We first show a dialogue between user and system in text form to the annotators and ask the annotators to label the user satisfaction of each user sentence at the *exchange-level*. We require annotators to rate user satisfaction based on past conversations, so the satisfaction is assessed before the user's sentence, not after writing the sentence. In this regard, we differ from previous annotation work [7, 44, 50]. The scale we asked annotators to follow was: (1) Very dissatisfied (the system fails to understand and fulfill user's request); (2) Dissatisfied (the system understands the request but fails to satisfy it in any way); (3) Normal (the system understands users request and either partially satisfies the request or provides information on how the request can be fulfilled); (4) Satisfied (the system understands and satisfies the user request, but provides more information than what the user requested or takes extra turns before meeting the request); and (5) Very satisfied (the system understands and satisfies the user request completely and efficiently).

Using a 5 point scale over a binary scale provides an option for the annotators to factor in their subjective interpretation of the extent of success or failure of a system's response to satisfy a user's request. In addition, we ask the annotators to rate the *dialogue-level* satisfaction to capture the overall satisfaction of a user's interaction with the system. We divide the data into two groups based on language, JDDC (Chinese) and Others (English). In each group, we randomly assign data to annotators to ensure that the different types of conversations in the group are evaluated according to a consistent standard. For the JDDC group, we also ask annotators to give a textual explanation for the rating.

Table 2: Statistics of the USS dataset.

Domain	JDDC	SGD	MultiWOZ	ReDial	CCPE
Language	Chinese	English	English	English	English
#Dialogues	3,300	1,000	1,000	1,000	500
Avg# Turns	32.3	26.7	23.1	22.5	24.9
#Utterances	54,517	13,833	12,553	11806	6,860
Rating 1	120	5	12	20	10
Rating 2	4,820	769	725	720	1,472
Rating 3	45,005	11,515	11,141	9,623	5,315
Rating 4	4,151	1,494	669	1,490	59
Rating 5	421	50	6	34	4

4.3 Measures and disclaimers

To guarantee annotation quality, we ask at least three annotators to repeatedly label the data. If there is a discrepancy among the three annotators (i.e., three annotators give three different ratings), we ask a fourth annotator to recheck it. We removed the results of annotators that were inconsistent with others. Finally, expert ratings are highly correlated with a Fleiss Kappa score of 0.574. See Table 2 for descriptive statistics of the USS dataset.

In all the provided instruction materials, we described the purpose of this data construction effort and pointed out that the data will only be used for research. We did not record any information about the annotators and warned the annotators not to divulge any of their private information.

5 EXPERIMENTS

5.1 Models used for comparison

Inspired by previous work [5, 23, 54], we consider three types of approach: Feature-based, RNN-based, and BERT.

5.1.1 Feature-based models. We use (1) TF-IDF, (2) the length of the last utterance (i.e., the number of words), and (3) position of the current utterance as the features in feature-based models. We compare several machine learning models that have popularly been used for text classification [1]: (1) logistic regression (LR), (2) support vector machines (SVM), and (3) XGBoost.

5.1.2 RNN-based models. Given the dialogue context $U = \{u_j\}_{j=1}^{t}$, we first encode it to get the context representation $\mathbf{h}^U$, and then predict the user satisfaction by $P(s \mid U) = \text{softmax}(\text{MLP}(\mathbf{h}^U))$. Inspired by previous work, we compare three methods for context representation encoding: (1) GRU, which first concatenates the dialogue history into a long sentence, and then feeds the sentence into a Bidirectional GRU (BiGRU) model. Then the context representation is defined as the average pooled outputs of the BiGRU model. (2) HiGRU, which explores the hierarchical structure. First, it encodes each utterance in the dialogue using a word-level BiGRU to get the utterance representations $\mathbf{h}^{u_j}$. Then it feeds the utterance representations into a sentence-level GRU, and define the context representation as the last hidden state of the sentence-level GRU [23]. (3) HiGRU+ATTN, which applies a two-level attention mechanism in HiGRU [54].

5.1.3 BERT-based model. Given the dialogue context $U = \{u_j\}_{j=1}^{t}$, we first concatenate it to a long sequence with $[SEP]$. Then we

encode it into a latent representation via BERT [14], and convert it into the condensed representation $\mathbf{h}^U$ through an average pooling operation. User satisfaction is predicted as $P(s \mid U) = \text{softmax}(\text{MLP}(\mathbf{h}^U))$.

5.2 Implementation details

To integrate the user satisfaction prediction and action prediction, we train two independent models for two tasks, in which action prediction takes the predicted output of satisfaction prediction model as the input. We use ground truth satisfaction in training and the model predicted satisfaction in testing. The Feature-based models are implemented using the scikit-learn toolkit. For the BERT-based model, we use BERT-Base (110M) pretrained weights[1] (hidden size is 768). We use the BERT vocabulary (size: 30,522) for all models (the Chinese BERT vocabulary for the JDDC domain), set the batch size = 64, the learning rate to 2e-5 for BERT and 1e-4 for others, use the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) to optimize parameters, use gradient clipping with a maximum gradient norm of 0.2, train up to 50 epochs on one NVIDIA TITAN RTX GPU, and select the best checkpoints based on performance on the validation set. Due to the serious imbalance of the satisfaction label, we up-sample the non-3 rating data during training. We take dialogue-level satisfaction as the last user utterance and use "overall" as the identification. As in previous work [9], we use 10-fold cross-validation to evaluate the outcome.

6 EVALUATION

6.1 Evaluation metrics

For the user satisfaction prediction task, following [44], we use the *Unweighted Average Recall* (UAR), the arithmetic average of all class-wise recalls, a linearly weighted version of *Cohen's Kappa*, and *Spearman's Rho* as evaluation metrics. We also use the *F1-score* for the *dissatisfactory* (rating < 3) class as the binary classification metric, as most turns and dialogues belong to the *satisfactory* (rating ≥ 3) class. For the user action prediction task, we use *Accuracy* (Acc, the proportion of predicted correct labels over the total number of predicted and actual labels for every utterance), *Precision* (Prec, the proportion of the predicted correct labels over the number of predicted labels), *Recall* (the proportion of the predicted correct labels over the number of actual labels), and the *F1-score* (the harmonic mean of precision and recall) as evaluation measures.

6.2 Experimental results

Table 3 shows the results for the user satisfaction prediction task. The best results in terms of the corresponding metric are shown in bold. If there are multiple similar best results, we show them all underlined. In general, HiGRU achieves the best overall performance (e.g., an absolute improvement of +3 for UAR, +2 for Kappa, and +4 for F1 over BERT in SGD data). BERT and HiGRU+ATTN can achieve performance comparable to HiGRU, followed by GRU. Among the 3 feature-based methods, SVM performs best, followed by LR. XGBoost is significantly weaker than other methods in all metrics, except Rho. Table 3 further shows that all deep learning methods perform better than feature-based metrics.

[1] https://github.com/huggingface/transformers

Table 4 shows the results for the user action prediction task. In general, the BERT-based model performs best among all methods, followed by HiGRU. BERT outperforms HiGRU on all performance measures except for the ReDial data, possibly due to the lack of sufficient training data. Among the 3 feature-based methods, XGBoost achieves the best performance, obtaining an absolute improvement of about +6 for Acc, +7 for Prec, +3 for Recall, and +4 for F1 compared to LR. XGBoost also outperforms GRU in many metrics.

6.3 Analysis

Since we have multiple domains in the dataset, we further analyze the cross-domain generalization capabilities of the user satisfaction prediction model. Table 5 shows the results. The rows and columns in Table 5 indicate training data and test data, respectively (e.g., 0.233 in the first column of the sixth row indicates that a BERT model trained on MultiWOZ can get a UAR score of 0.233 on SGD data). In terms of datasets, the models trained on SGD and MultiWOZ get the best performance on each other's data respectively, and the models trained on ReDial get the best performance on CCPE, possibly due to the similarity between domains. The model trained on CCPE has relatively poor generalization ability, possibly due to limited training data size. In terms of methods, BERT achieves better generalization performance than SVM and HiGRU, possibly due to the improvement of pre-training on the large-scale corpus.

7 UTILIZATION OF THIS RESOURCE

We have developed resources that are meant to help answer the question of what is a good dialogue. Our annotations and prediction task offer a better characterization of what is a good dialogue than existing datasets. Exchange-level user satisfaction and action prediction can reflect what kind of system behavior will bring positive user satisfaction and what behavior will harm the user experience, which makes our method applicable to many related fields.

7.1 Building human-like user simulation

In most prior work, user simulations mechanically give the slots, and thus measure very limited aspects of a dialogue. Building a human-like user simulation remains an open challenge. In this study, we propose the task of user satisfaction simulation and release a dataset for the task. Inspired by previous work on similar tasks [5, 23, 54], we provide a series of baselines. However, due to the challenging nature of the task, there is plenty of room to improve user satisfaction prediction, and to explore how user satisfaction prediction can be combined with action prediction. Response generation based on user satisfaction (i.e., reflect user satisfaction in a generated utterance) is still an open problem. Previous work on open-domain dialogue may serve as a reference [58]. In addition to user satisfaction, how to ground a user simulator by introducing external knowledge [35, 36, 47, 52] and persona [32] to establish a more human-like user simulator has not yet been studied.

7.2 Future applications

The USS dataset can be used not only for user simulation but also for other conversational information access tasks. As a user satisfaction annotation dataset that exceeds existing ones in scale, our data can

Table 3: Performance for user satisfaction prediction. Bold face indicates the best result in terms of the corresponding metric. Underline indicates comparable results to the best one.

Domain	JDDC				SGD				MultiWOZ				ReDial				CCPE			
	UAR	Kappa	Rho	F1	UAR	Kappa	Rho	F1	UAR	Kappa	Rho	F1	UAR	Kappa	Rho	F1	UAR	Kappa	Rho	F1
LR	0.221	0.054	0.400	0.011	0.211	0.049	0.251	0.005	0.214	0.042	0.599	0.009	0.211	0.040	0.240	0.008	0.214	0.060	0.669	0.025
SVM	0.235	0.061	0.347	0.026	0.230	0.074	0.169	0.020	0.215	0.030	0.425	0.021	0.209	0.038	0.205	0.015	0.212	0.027	0.534	0.040
XGBoost	0.205	0.007	**0.584**	0.003	0.202	0.011	0.442	0.001	0.200	0.002	0.690	0.001	0.207	0.030	0.391	0.002	0.200	0.001	0.707	0.004
HiGRU+ATTN	0.330	0.115	0.502	0.180	0.262	0.082	0.475	0.058	0.224	0.142	0.842	0.197	**0.261**	0.097	**0.441**	0.118	0.223	0.109	0.869	0.214
HiGRU	**0.339**	0.126	0.524	0.171	**0.293**	**0.118**	0.451	**0.086**	0.225	**0.143**	**0.886**	**0.238**	0.257	0.084	0.324	0.083	**0.237**	**0.167**	0.881	**0.274**
GRU	0.302	0.092	0.497	0.132	0.245	0.072	0.248	0.027	0.231	0.105	0.813	0.167	0.254	0.104	0.421	0.121	0.226	0.124	0.880	0.207
BERT	0.329	**0.131**	0.554	**0.185**	0.261	0.094	**0.477**	0.048	**0.256**	0.133	0.823	0.224	0.257	**0.122**	0.390	**0.125**	0.232	0.147	**0.891**	0.245

Table 4: Performance for user action prediction. Bold face indicates the best result in terms of the corresponding metric. Underline indicates comparable results to the best one.

Domain	JDDC				SGD				MultiWOZ				ReDial				CCPE			
	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
LR	0.565	0.208	0.123	0.133	0.460	0.321	0.308	0.309	0.414	0.150	0.130	0.134	0.495	0.467	0.472	0.464	0.509	0.325	0.314	0.316
SVM	0.493	0.214	0.139	0.147	0.451	0.344	0.351	0.345	0.374	0.141	0.138	0.135	0.459	0.423	0.444	0.427	0.462	0.327	0.327	0.322
XGBoost	0.621	0.270	0.138	0.165	0.516	0.395	0.370	0.370	0.479	0.226	0.126	0.139	0.593	0.540	0.509	0.506	0.553	0.380	0.349	0.356
HiGRU+ATTN	**0.623**	0.363	0.176	0.194	0.617	0.498	0.481	0.481	0.487	0.221	0.152	0.155	0.590	0.548	0.512	0.488	0.611	0.421	0.408	0.411
HiGRU	0.618	0.370	0.196	**0.229**	0.643	0.534	0.505	0.507	0.518	0.216	0.162	0.167	**0.622**	**0.584**	**0.532**	**0.534**	0.672	0.503	0.472	0.482
GRU	0.598	0.337	0.166	0.187	0.444	0.322	0.304	0.298	0.460	0.211	0.124	0.129	0.599	0.536	0.494	0.457	0.545	0.550	0.354	0.354
BERT	0.614	**0.391**	**0.199**	0.224	**0.661**	**0.570**	**0.572**	**0.560**	**0.519**	**0.255**	**0.183**	**0.191**	0.614	0.573	0.531	0.530	**0.674**	**0.696**	**0.495**	**0.496**

Table 5: Cross-domain performance for user satisfaction prediction. Report UAR.

From	To	SGD	MWOZ	ReDial	CCPE
SGD	SVM	0.230	0.209	0.211	0.198
	HiGRU	0.293	0.240	0.230	0.212
	BERT	0.261	**0.249**	**0.254**	0.223
MWOZ	SVM	0.208	0.215	0.206	0.208
	HiGRU	0.224	0.225	0.221	0.219
	BERT	**0.233**	0.256	0.219	0.226
ReDial	SVM	0.216	0.227	0.221	0.199
	HiGRU	0.211	0.221	0.261	0.220
	BERT	0.228	0.218	0.257	**0.239**
CCPE	SVM	0.217	0.208	0.218	0.214
	HiGRU	0.211	0.223	0.227	0.237
	BERT	0.216	0.213	0.219	0.232

facilitate research on user satisfaction modeling [39] and POMDP-based dialogue systems [30, 56]. Moreover, the USS dataset can also facilitate research into dialogue breakdown detection, and human-machine hand-off prediction [34]. In the JDDC domain, we provide annotators' explanations on user satisfaction annotations, which includes a total of 9,900 explanation texts. This information can be applied to user studies of user satisfaction, and interpretability studies of evaluations.

8 CONCLUSION

We have proposed the task of simulating user satisfaction for evaluating task-oriented dialogue systems, so as to enhance the evaluation of dialogue systems. We have collected and released a new benchmark dataset, namely USS, for the proposed task. Our dataset contains a total of 6,800 dialogues spanning multiple domains. We have introduced three baselines for our task: feature-based, RNN-based, and BERT-based methods. Experiments conducted on the newly collected dataset suggest that distributed representations do outperform feature-based methods. Besides, HiGRU achieves the best performance in in-domain user satisfaction prediction, while a BERT-based method has better cross-domain generalization ability.

As to our future work, we would like to continue to investigate the combination of the user satisfaction prediction and action prediction task, and response generation based on user satisfaction.

DATA

We share the USS dataset at https://github.com/sunnweiwei/user-satisfaction-simulation.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China with grant No. 2020YFB1406704, the Natural Science Foundation of China (61972234, 61902219, 62072279), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), the Tencent WeChat Rhino-Bird Focused Research Program (JR-WXG-2021411), the Fundamental Research Funds of Shandong University, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Charu Aggarwal and ChengXiang Zhai. 2012. Mining Text Data. In *Springer US*.

[2] David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. *Luke, I am Your Father: Dealing with Out-of-Domain Requests by Using Movies Subtitles*. Springer International Publishing.

[3] Layla El Asri, Jing He, and Kaheer Suleman. 2016. A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems. In *INTERSPEECH*.

[4] Rafael E. Banchs and Haizhou Li. 2013. IRIS: a Chat-Oriented Dialogue System Based on the Vector Space Model. In *ACL*.

[5] Lina M. Rojas Barahona. 2021. Is the User Enjoying the Conversation? A Case Study on the Impact on the Reward Function. *ArXiv* abs/2101.05004 (2021).

[6] Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, B. Thomson, Jason Williams, Kai Yu, Steve Young, and Maxine Eskénazi. 2011. Spoken Dialog Challenge 2010: Comparison of Live and Control Test Results. In *SIGDIAL Conference*.

[7] Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019. Multi-domain Conversation Quality Evaluation via User Satisfaction Estimation. *ArXiv* abs/1911.08567 (2019).

[8] Praveen Kumar Bodigutla, Aditya Tiwari, Josep Valls-Vargas, Lazaros Polymenakos, and Spyridon Matsoukas. 2020. Joint Turn and Dialogue level User Satisfaction Estimation on Multi-Domain Conversations. *ArXiv* abs/2010.02495 (2020).

[9] Wanling Cai and Li Chen. 2020. Predicting User Intents and Satisfaction with Dialogue-based Conversational Recommendations. *ACM UMAP* (2020).

[10] Hongshen Chen, Zhaochun Ren, Jiliang Tang, Eric Yihong Zhao, and Dawei Yin. 2018. Hierarchical Variational Memory Network for Dialogue Generation. In *WWW*. 1653–1662.

[11] Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service. In *LREC*.

[12] Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. 2005. Human-Computer Dialogue Simulation using Hidden Markov Models. *IEEE Workshop on ASRU, 2005.* (2005), 290–295.

[13] Jan Deriu, A. Rodrigo, Arantxa Otegi, Guillermo Echegoyen, S. Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review* 54 (2020), 755 – 810.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[15] Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. User Modeling for Spoken Dialogue System Evaluation. *IEEE Workshop on ASRU* (1997), 80–87.

[16] Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling User Satisfaction with Hidden Markov Models. In *SIGDIAL*.

[17] Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Anuj Kumar Goyal, Peter Ku, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *LREC*.

[18] Kallirroi Georgila, James Henderson, and Oliver Lemon. 2005. Learning User Simulations for Information State Update Dialogue Systems. In *INTERSPEECH*.

[19] Sunao Hara, Norihide Kitaoka, and Kazuya Takeda. 2010. Estimation Method of User Satisfaction Using N-gram-based Dialog History Model for Spoken Dialog System. In *LREC*.

[20] Homa B Hashemi, Amir Asiaee, and Reiner Kraft. 2016. Query Intent Detection using Convolutional Neural Networks. In *WSDM, Workshop on Query Understanding*.

[21] Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Issues in Predicting User Satisfaction Transitions in Dialogues: Individual Differences, Evaluation Criteria, and Prediction Models. In *IWSDS*.

[22] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li'e Chen. 2020. A Survey on Conversational Recommender Systems. *ArXiv* abs/2004.00646 (2020).

[23] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. HiGRU: Hierarchical Gated Recurrent Units for Utterance-level Emotion Recognition. *NAACL-HLT* (2019), 397–406.

[24] Xisen Jin, Wenqiang Lei, Zhaochun Ren, Hongshen Chen, Shangsong Liang, Yihong Zhao, and Dawei Yin. 2018. Explicit State Tracking with Semi-Supervision for Neural Dialogue Generation. In *ACM CIKM*. 1403–1412.

[25] Filip Jurcícek, Simon Keizer, Milica Gasic, François Mairesse, B. Thomson, Kai Yu, and S. Young. 2011. Real User Evaluation of Spoken Dialogue Systems Using Amazon Mechanical Turk. In *INTERSPEECH*.

[26] Florian Kreyssig, Iñigo Casanueva, Pawel Budzianowski, and Milica Gasic. 2018. Neural User Simulation for Corpus-based Policy Optimisation for Spoken Dialogue Systems. *ArXiv* abs/1805.06966 (2018).

[27] Lori Lamel, Sophie Rosset, Jean-Luc Gauvain, Samir Bennacef, Martine Garnier-Rizet, and Bernard Prouts. 2000. The LIMSI ARISE system. *Speech Commun.* 31, 4 (2000), 339–353.

[28] 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 *ACL*. 1437–1447.

[29] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *SIGKDD*.

[30] Oliver Lemon and Olivier Pietquin. 2012. Data-Driven Methods for Adaptive Spoken Dialogue Systems. In *Springer New York*.

[31] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL*.

[32] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and W. Dolan. 2016. A Persona-Based Neural Conversation Model. *ACL*.

[33] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. *ArXiv* abs/1812.07617 (2018).

[34] Jiawei Liu, Zhe Gao, Yangyang Kang, Zhuoren Jiang, Guoxiu He, Changlong Sun, Xiaozhong Liu, and Wei Lu. 2021. Time to Transfer: Predicting and Evaluating Machine-Human Chatting Handoff. *AAAI* (2021).

[35] Longxuan Ma, Weinan Zhang, Runxin Sun, and Ting Liu. 2020. A Compare Aggregate Transformer for Understanding Document-grounded Dialogue. In *Findings of EMNLP*. 1358–1367.

[36] Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Z. Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. DukeNet: A Dual Knowledge Interaction Network for Knowledge-Grounded Conversation. *SIGIR*.

[37] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain Dialog State Tracking using Recurrent Neural Networks. In *IJCNLP (Volume 2: Short Papers)*. 794–799.

[38] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *ACL*. 1777–1788.

[39] Louisa Pragst, Stefan Ultes, and Wolfgang Minker. 2016. Recurrent Neural Network Interaction Quality Estimation. In *IWSDS*.

[40] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *SIGdial*.

[41] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *AAAI*.

[42] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In *HLT-NAACL*.

[43] Konrad Scheffler and Steve Young. 2000. Probabilistic Simulation of Human-Machine Dialogues. *IEEE ICASSP (Cat. No.00CH37100)* 2 (2000), II1217–II1220 vol.2.

[44] Alexander Schmitt and Stefan Ultes. 2015. Interaction Quality: Assessing the Quality of Ongoing Spoken Dialog Interaction by Experts - And How it Relates to User satisfaction. *Speech Commun.* 74 (2015), 12–36.

[45] Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System. In *LREC*.

[46] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*.

[47] Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Conversations Powered by Cross-Lingual Knowledge. In *SIGIR*.

[48] Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2013. On Quality Ratings for Spoken Dialogue Systems - Experts vs. Users. In *HLT-NAACL*.

[49] Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. In *ICML*.

[50] Marilyn A. Walker, Diane Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *ArXiv* cmp-lg/9704004 (1997).

[51] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *EACL*. 438–449.

[52] Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *ACL*. 1835–1845.

[53] Zhao Yan, Nan Duan, Peng Chen, M. Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. In *AAAI*.

[54] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *HLT-NAACL*.

[55] Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. 2013. Recurrent Neural Networks for Language Understanding. In *Interspeech*.

[56] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based Statistical Spoken Dialog Systems: A Review. *Proc. IEEE* 101, 5 (2013).
[57] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. *ACM SIGKDD* (2020).
[58] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *AAAI*.