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

Wewei Sun¹* Shuo Zhang²* Krisztian Balog³ Zhaochun Ren¹† Pengjie Ren¹ Zhumin Chen¹ Maarten de Rijke⁴,⁵

¹Shandong University, Qingdao, China ²Bloomberg, London, United Kingdom ³University of Stavanger, Stavanger, Norway ⁴University of Amsterdam ⁵Ahold Delhaize Research, Amsterdam, The Netherlands

sunweiwei@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 evaluation 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).

KEYWORDS

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

∗Equal contribution. †Corresponding author.

ACM Reference Format:

Wewei 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: (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 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 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] 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 Cuayahuitl 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 user’s failure to understand the system’s feedback 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 user’s 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 \( \mathcal{D} = \{(U_i, a_i, s_j)\}_{i=1}^N \), where for all \( i \in [1, N] \), \( U_i \) is the dialogue context, \( a_i \) is the next-turn user action, and \( s_j \) 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 is 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) [35], 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).

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     |
(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.

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 \( h_U \), and then predict the user satisfaction by \( P(s \mid U) = \text{softmax}(\text{MLP}(h_U^T)) \).

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 \( h^{\text{U}} \). 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

| 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 | 11,806 | 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.
encode it into a latent representation via BERT [14], and convert it into the condensed representation \( h^U \) through an average pooling operation. User satisfaction is predicted as \( P(s \mid U) = \text{softmax}(\text{MLP}(h^U)) \).

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 (\( \beta_1 = 0.9, \beta_2 = 0.999, \text{and} \ e = 10^{-8} \)) 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 achieves better generalization performance than SVM and HiGRU, possibly due to limited training data size. In terms of methods, BERT performs better than feature-based metrics, followed by HiGRU. 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. 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

\(^{1}\text{https://github.com/huggingface/transformers}\)
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 | 0.111 | 0.211 | 0.049 | 0.251 | 0.005 | 0.214 | 0.042 | 0.359 | 0.009 | 0.211 | 0.040 | 0.240 | 0.008 | 0.214 | 0.060 | 0.669 | 0.025 |
| SVM    | 0.235 | 0.661 | 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 |

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 | 0.533 | 0.380 | 0.349 | 0.356 |

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

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

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.
[56] Steve Young, Milica Gasić, 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.