Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 726–734
Marseille, 11–16 May 2020

Predicting Ratings of Real Dialogue Participants from Artificial Data and Ratings of Human Dialogue Observers

Kallirroi Georgila, Carla Gordon, Volodymyr Yanov, David Traum
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Los Angeles, CA 90094-2536, USA
{kgeorgila, cgordon, yanov, traum}@ict.usc.edu

Abstract

We collected a corpus of dialogues in a Wizard of Oz (WOz) setting in the Internet of Things (IoT) domain. We asked users participating in these dialogues to rate the system on a number of aspects, namely, intelligence, naturalness, personality, friendliness, their enjoyment, overall quality, and whether they would recommend the system to others. Then we asked dialogue observers, i.e., Amazon Mechanical Turkers (MTurkers), to rate these dialogues on the same aspects. We also generated simulated dialogues between dialogue policies and simulated users and asked MTurkers to rate them again on the same aspects. Using linear regression, we developed dialogue evaluation functions based on features from the simulated dialogues and the MTurkers' ratings, the WOz dialogues and the MTurkers' ratings, and the WOz dialogues and the WOz participants' ratings. We applied all these dialogue evaluation functions to a held-out portion of our WOz dialogues, and we report results on the predictive power of these different types of dialogue evaluation functions. Our results suggest that for three conversational aspects (intelligence, naturalness, overall quality) just training evaluation functions on simulated data could be sufficient.

Keywords: dialogue evaluation functions, real and simulated dialogues, Internet of Things

1. Introduction

In order to build a dialogue system for a particular domain, data in this domain are required. Ideally these data should record interactions between real users and a dialogue system, or, if a dialogue system is not available (which is very common in the initial stages of development), interactions between real users and a Wizard (a human playing the role of the system), in a so called Wizard of Oz (WOz) setting (Dahlbäck et al., 1993). However, this approach can be quite expensive and time consuming. Building the initial dialogue system or the WOz environment can be costly. Furthermore, training one or more Wizards and recruiting human participants to interact with the Wizards can significantly add to the overall cost, and the whole process can be time consuming. An alternative cheaper approach is to generate artificial data, either by having dialogue policies interact with simulated users and generate a very large number of dialogues (El Asri et al., 2016), or by having linguists write a variety of dialogues (Georgila et al., 2018).

Here we use human-system data collected in a WOz setting, and simulated dialogues generated by having dialogue policies interact with simulated users. Simulated dialogues can be very useful in bootstrapping the development of a dialogue system, but they do not fully represent how real users would react in real time to system interventions in a specific context. Real user behavior can be unpredictable and vary widely from user to user. Thus it is difficult for simulated dialogues to capture all aspects of real use, especially the relative frequencies of different kinds of issues.

Our goal is to develop dialogue quality evaluation functions to predict ratings of real users interacting with a dialogue system. We hypothesize that dialogue quality evaluation functions trained on real users' dialogues and real users' ratings (i.e., ratings of human users participating in the dialogue) would be more highly predictive of real users' ratings (on unseen data) than evaluation functions trained on real users' dialogues and human observers' ratings (i.e., ratings of Amazon Mechanical Turkers who read the real dialogues and rate them), or evaluation functions trained on simulated dialogues and human observers' ratings (i.e., ratings of Amazon Mechanical Turkers who read the simulated dialogues and rate them). The question that we want to address is what kind of error in our predictions we should expect when we use simulated data and ratings of Amazon Mechanical Turkers (MTurkers), or real users' data and MTurkers' ratings to train our evaluation functions, instead of the gold-standard of having access to both real users' dialogues and real users' ratings. When deploying dialogue systems typically we collect dialogues from real users but not their ratings, as it would be very disruptive to the users if we constantly asked them to evaluate their interaction with the system. Thus having MTurkers rate real dialogues is a realistic approach.

We collected a corpus of dialogues in a WOz setting in the Internet of Things (IoT) domain. The IoT is the network of physical devices (e.g., home appliances, health monitoring devices, etc.) connected to the Internet. IoT devices can be controlled each one separately by individual apps, or all together via an integrated app. Alternatively, IoT devices can be controlled by a smart assistant via human-system dialogue interaction (Jeon et al., 2016), which is the approach that we follow here. We asked the users participating in these dialogues to rate the system on a number of aspects, namely, intelligence, naturalness, personality, friendliness, their enjoyment, overall quality, and whether they would recommend the system to others. Then we asked MTurkers to rate these dialogues on the same aspects. We also generated simulated dialogues between dialogue policies and simulated users and again asked MTurkers to rate them on the same aspects. Using linear regression, we developed dialogue evaluation functions based on features from the simulated dialogues and the MTurkers' ratings, the WOz dia-

logues and the MTurkers' ratings, and the WOz dialogues and the WOz participants' (real users') ratings. We applied all these dialogue evaluation functions to a held-out portion of our WOz dialogues, and we report results on the predictive power of these different types of dialogue evaluation functions.

To our knowledge no such study has been performed before, and certainly not in the IoT domain. Previous work on dialogue evaluation in the IoT domain considered dialogue quality evaluation functions trained on dialogues written by linguists and MTurkers' ratings (Georgila et al., 2018). Furthermore, Gordon et al. (2018) established links between objective measures and more nuanced subjective judgements, namely, intelligence, personality, pleasantness, and naturalness, also in the IoT domain.

2. Related Work

Hastie (2012) presents an overview of evaluation frameworks and metrics that have been proposed in the literature for measuring the quality of human-system dialogue interaction, mainly for task-oriented dialogue systems. Some of these metrics are subjective (e.g., user satisfaction, perceived task completion, etc.), while others are objective (e.g., word error rate, dialogue length, etc.). Objective measures can be calculated from the interaction logs while subjective assessments can be collected via surveys and questionnaires (Hone and Graham, 2000; Paksima et al., 2009). PARADISE is perhaps the most well-known framework for evaluating dialogue systems, and an attempt to automate the evaluation process (Walker et al., 2000). PARADISE seeks to optimize a desired quality such as user satisfaction by formulating it as a linear combination of a variety of metrics, such as task success and dialogue cost (e.g., dialogue length, speech recognition errors, etc.). The contribution of each factor is determined by weights calculated via linear regression. The advantage of this method is that once a desired quality has been formulated as a realistic evaluation function, it can be optimized by controlling the factors that affect it. In the example above, user satisfaction can be optimized by increasing task success, and minimizing dialogue length and speech recognition errors.

Reinforcement learning (RL) is a very popular approach to learning dialogue policies from data or simulated users (SUs) (Jurčíček et al., 2012). In RL, a typical reward function is for the system to earn a number of points for a fully or partially successful dialogue, and subtract a penalty per system turn to ensure that the learned dialogue policies will not favor lengthy and tedious dialogues (Henderson et al., 2008). Note however that longer dialogue lengths are not necessarily indicative of poor dialogue quality but depending on the task they may actually indicate user engagement and satisfaction (Foster et al., 2009).

Schatzmann et al. (2006) present an overview of metrics that have been proposed in the literature for measuring the quality of SUs used for training and evaluating dialogue policies. The action generated by the SU is compared against the user action in a human-human or human-system reference corpus (in the same dialogue context), and measures such as precision, recall, accuracy, and perplexity are used (Schatzmann et al., 2005; Georgila et al., 2005; Georgila et al., 2006; Pietquin and Hastie, 2013). Also, to take into account the fact that SU actions are generated based on a probability distribution, expected precision, expected recall, and expected accuracy are used (Georgila et al., 2006). However, these metrics can be problematic because if a SU action is not the same as the user action in the reference corpus, this does not necessarily mean that it is a poor action. Also, once a user or system response deviates from the corresponding action in the reference corpus, the remaining dialogue will unfold in an entirely different way than the fixed dialogue in the reference corpus, which will make further comparisons meaningless.

In non-task-oriented dialogue systems (e.g., chatbots) developing robust evaluation metrics can be even harder than for task-oriented dialogue (Misu et al., 2012). Here it is not clear what success means and task-specific objective metrics are not appropriate. Instead subjective evaluations for appropriateness of responses can be much more meaningful, which has led to the development of coding schemes for response appropriateness in such cases (Traum et al., 2004; Robinson et al., 2010).

Currently, word-overlap similarity metrics such as BLEU, METEOR, and ROUGE (originally employed in machine translation and summarization) are widely used for measuring chatbot dialogue quality. However, BLEU, METEOR, and ROUGE suffer from the same problems as the aforementioned SU evaluation metrics. In fact it has been shown that BLEU, METEOR, and ROUGE do not correlate well with human judgements of dialogue quality (Liu et al., 2016). Discriminative BLEU, a variation of BLEU where reference strings are scored for quality by human raters, was found to correlate better with human judgements than standard BLEU (Galley et al., 2015). To address the issues with BLEU, METEOR, and ROUGE, next utterance classification was introduced as a method for evaluating chatbots (Lowe et al., 2016), but the proposed metric recall@k is very similar to the recall metric previously used for evaluating SUs, and consequently has the same limitations. Also, topic-based metrics for chatbot evaluation (topic breadth and topic depth) were found to correlate well with human judgements (Guo et al., 2017).

3. Wizard of Oz Data Collection

We built a GUI-based environment for our Wizard of Oz (WOz) data collection (Gordon et al., 2019). A human Wizard plays the role of the system by pressing buttons in a GUI. Each button corresponds to a Wizard action which is then transformed into a sentence (through template-based natural language generation) and provided to the user. Thus the human user receives system output and responds accordingly. Note that there are additional buttons on the Wizard's GUI that do not send messages to the user, but are used for updating the status of the devices. Both the user and the Wizard can see the dialogue as it unfolds. They can also see the status of the devices and the changes occurring during the interaction. This is particularly important because how users perceive the status of the devices can affect the flow of the interaction. For example, if the system confirms a request to move the channel of the speaker up one and the user then changes their mind to ask for moving

Dialogue 1
no misunderstandings, direct, explicit
Task: AC wifi in kitchen, turn on.
(AC wifi in kitchen: status is off)
U: Turn on the wifi for the AC in the kitchen.
S: Ok. I connected the wifi for the AC in the kitchen.
(AC wifi in kitchen: status is on)
Dialogue 2
misunderstandings, conversational, implicit
Task: light in bedroom, turn on in 15 min.
(Light in bedroom: status is off)
U: In 15 min, turn on the light in the bedroom.
S: Sure thing! I'll take care of that in 5 min.
(Light in bedroom: status scheduled to become on in 5 min)
U: Not 5 minutes. It should be on in 15 min.
S: Sorry! I took care of that for you.
(Light in bedroom: status scheduled to become on in 15 min)
Dialogue 3
misunderstandings, direct, implicit
Task: speaker in living room, turn off;
TV in living room, turn off;
speaker in bedroom, turn on.
(Speaker in living room: status is on)
(TV in living room: status is on)
(Speaker in bedroom: status is off)
U: Turn off the speaker and the TV in the living room.
S: Ok. I did that.
(Speaker in living room: status is off)
(TV in living room: status is off)
U: Turn on the speaker in the bedroom.
S: You would like the speaker on in the living room?
U: I would like the speaker on in the bedroom.
S: Ok. It has been done.
(Speaker in bedroom: status is on)

Figure 1: Example WOz dialogues.

the channel one down, correct system behavior will depend on whether the system has already performed the action at the time of correction (if so, then the speaker might need to move down 2 channels, if not, only one). Another example of the importance of showing the status of the devices is that the user could ask the system to perform an action and the system could respond "done". Unless the user can see the status of the devices, there is no way they can be sure that the action was actually performed.

The devices that we consider in our IoT ontology are washer, speaker, bulb, TV, and air-conditioner (AC). We may also have multiple devices of the same type (e.g., there can be a TV in the bedroom, a TV in the living room, and a TV in the kitchen). We created 8 system profiles representing a mixture of 3 different system features:

- Conversational behavior (e.g., "thanks, it's my pleasure to help") vs. formal/direct behavior (e.g., "thank you").
- Implicit system responses (e.g., "your request has been taken care of") vs. explicit system responses (e.g., "I've turned on the light in the kitchen").
- Errors (misunderstandings) vs. no errors: Error conditions represent all possible errors a Virtual Home Assistant might make, including carrying out a request for the wrong device, for the device in the wrong location, for the wrong time (e.g., in 5 minutes instead of in 15 minutes), turning a device or device component on instead of off, or vice versa, and changing settings incorrectly (e.g., setting volume to 8 instead of 4, or setting temperature to hot instead of cold).

The features of the profile serve as a guide for the Wizard. Thus the Wizard uses buttons that generate appropriate language for conversational vs. formal/direct behavior and explicit vs. implicit system responses. Also, if the profile says that there should be misunderstandings, the Wizard pretends not to understand the user's request and deliberately confirms the wrong command.

Figure 1 shows three WOz dialogues. There are no errors in the first dialogue, whereas the second and third dialogues contain misunderstandings. The first and third ones have a direct style, while the second is conversational, and finally the first one has explicit acknowledgement of what the system will do, while the second and third just confirm receipt and performance without specifying what was done.

We collected data from 18 real users in our WOz setting. Each user interacted with 4 system profiles. We designed 36 tasks and each user was required to complete 12 tasks (randomly selected from the 36 tasks), 3 tasks per system profile. The tasks were representative of typical tasks one might accomplish with the help of a Virtual Home Assistant, such as turning on the lights, turning off the TV, etc. There was a set of 6 tasks specific to each of the 5 devices (TV, bulb, speaker, washer, AC) as well as a set of 6 tasks that include multiple device types (TV/washer, bulb/TV/speaker, etc.). Each task had an error condition which was determined by which system profile was paired with the task. If the profile included errors, the error condition would be carried out, otherwise it would not. There was a post-dialogue survey at the end of each completed task where the WOz participants were asked to rate on a 7-point Likert scale (1:low, 7:high) the system in each one of their interactions regarding the following aspects: intelligence, naturalness, personality, friendliness, their enjoyment, overall quality, and whether they would recommend the system to others. Note that the users had no information about the profiles and different possible behaviors of the system (Wizard). They were just instructed to interact with the system and rate their interaction based on their own interpretation of the rated aspects.

The next step was to collect MTurkers' ratings for the WOz dialogues. We recruited 101 MTurkers. Each MTurker was asked to rate 10 individual WOz dialogues (randomly selected from our corpus of WOz dialogues), and in particular the system in terms of intelligence, naturalness, personality, friendliness, their enjoyment, overall quality, and whether they would recommend the system to others. Basically the MTurkers had to answer the same survey questions as the real users in the WOz data collection, and similarly to the

Real Users' Rating	Feature	Pearson's r
Intelligence	Misund	-0.65***
Naturalness	Convers	0.28***
Naturalness	Misund	-0.31***
Overall quality	Misund	-0.58***
Recommend	Misund	-0.54***
Personality	Convers	0.41***
Personality	Misund	-0.17*
Enjoyment	Convers	0.18*
Enjoyment	Misund	-0.49***
Friendliness	Convers	0.32***
Friendliness	Misund	-0.23**

Table 1: Pearson's r correlations of real users' ratings with dialogue features, WOz dialogues (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, with Holm-Bonferroni correction).

MTurkers' Rating	Feature	Pearson's r
Intelligence	Misund	-0.50***
Naturalness	Convers	0.54***
Overall quality	Convers	0.23**
Overall quality	Misund	-0.50***
Recommend	Convers	0.21***
Recommend	Misund	-0.45***
Personality	Convers	0.48***
Enjoyment	Convers	0.34***
Enjoyment	Misund	-0.31***
Friendliness	Convers	0.36***
Friendliness	Misund	-0.23**

Table 2: Pearson's r correlations of MTurkers' ratings with dialogue features, WOz dialogues (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, with Holm-Bonferroni correction).

real users they had no information about the profiles and the different possible behaviors of the system (Wizard). They were just instructed to rate the dialogues based on their own interpretation of the rated aspects.

Due to software issues we ended up with 186 WOz dialogues where we had both WOz participants' ratings and MTurkers' ratings. Because of how we had set up the system profiles, for each one of these WOz dialogues we also have the following 3 binary features: conversational, explicit, and misunderstandings. Thus, for the conversational system profile, conversational=1. For the direct system profile, conversational=0. For the explicit system responses profile, explicit=1. For the implicit system responses profile, explicit=0. When there are errors in a dialogue, misunderstandings=1, otherwise misunderstandings=0.

We calculated pairwise Pearson correlations between the features in our WOz data and real users' ratings. Note that "***" means that the correlation is very significant ($p < 0.001$), "**" means that the correlation is significant ($p < 0.01$), and "*" means that the correlation is borderline significant ($p < 0.05$); the p values are corrected for multiple comparisons using the method of Holm-Bonferroni. We found that overall quality and recommend system to others have high correlations with intelligence (0.88*** and 0.85*** respectively). Also, recommend system to others and enjoyment have high correlations with overall quality (0.93*** and 0.87*** respectively). Enjoyment has a high correlation with recommend system to others (0.90***). Table 1 shows the correlations of the real users' ratings with the conversational and misunderstandings binary features in the WOz dialogues. Correlations with the explicit feature were not significant.

We also calculated pairwise Pearson correlations between the features in our WOz data and MTurkers' ratings. Similarly to the real users' ratings, overall quality and recommend system to others have high correlations with intelligence (0.81*** and 0.86*** respectively). Recommend system to others has a high correlation with overall quality (0.85***). Personality has a high correlation with naturalness (0.88***), and enjoyment has a high correlation with recommend system to others (0.86***). Table 2 shows the correlations of the MTurkers' ratings with the conversational and misunderstandings binary features in the WOz dialogues. Correlations with the explicit feature were not significant.

Real Users' Rating	MTurkers' Rating	Pearson's r
Intelligence	Intelligence	0.39***
Naturalness	Naturalness	0.20**
Overall quality	Overall quality	0.43***
Recommend	Recommend	0.40***
Personality	Personality	0.27***
Enjoyment	Enjoyment	0.25***
Friendliness	Friendliness	0.27***

Table 3: Pearson's r correlations of real users' ratings with MTurkers' ratings, WOz dialogues (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, with Holm-Bonferroni correction).

In both Table 1 and Table 2 we can see positive correlations of ratings with conversational system behavior, and negative correlations of ratings with misunderstandings. It is interesting that while the real users and the MTurkers were rating the same dialogues the resulting correlations are different. This shows that how dialogue participants perceive their interaction with the system is different from how dialogue observers perceive the interaction, which of course is not surprising. Table 3 shows the correlations between real users' ratings and MTurkers' ratings for the same aspect, e.g., correlation between naturalness and naturalness, intelligence and intelligence, etc. The correlations are positive and significant but not that high, which again is not surprising.

4. Simulated Data

We developed an agenda-based simulated user (SU) (Schatzmann and Young, 2009) for the IoT domain. We also developed a system policy that interacts with the SU. The agenda can be thought of as a stack containing the SU's pending actions, also called speech acts, which are required for accomplishing the SU's goal. For example, the agenda

could be initialized with requests for changing the status of devices. Based on hand-crafted rules (informed by our data), new speech acts are generated and pushed onto the agenda as a response to the system's actions. For example, if the system says "please specify the location of the device", a speech act for providing the location of the device will be pushed onto the agenda. At the same time, no longer relevant speech acts will be removed from the agenda. When the SU is ready to respond, 1, 2, or more speech acts will be popped off the agenda based on a random probability distribution. The generated simulated dialogues also include the status of the devices and their updates as the dialogue unfolds.

This is a very sophisticated SU that can handle multiple devices, misunderstandings, and different user types (conversational vs. direct/formal; preferring explicit vs. implicit system responses). The tasks, status of devices, and different user types are randomly initialized at the beginning of each dialogue. Also, the system policy that we developed can exhibit different kinds of behavior similar to the WOz setting (misunderstandings vs. no misunderstandings, conversational vs. direct system behavior, implicit vs. explicit system behavior), and these types of behaviors are randomly initialized at the beginning of each dialogue. The natural language generation part of the system policy uses different templates for misunderstandings vs. no misunderstandings, conversational vs. direct system behavior, and implicit vs. explicit system behavior.

There are 3 types of SUs: (1) Partial-input: the SU does not always provide the full command, e.g., "turn on the light" (here the location is missing). The decisions on what to generate are made randomly during the interaction. This SU can deal with misunderstandings, that is, it can understand that the system got confused and correct the system or reprovide the request. (2) Full-input: the SU always provides the full command, e.g., "turn on the light in the kitchen". This SU can also deal with misunderstandings. (3) Baseline: the SU is full-input and is not able to deal with misunderstandings. This SU always provides the full user request, regardless of how the conversation unfolds and even if this request has been partially fulfilled, and it repeats this request verbatim until all tasks have been completed, or after a pre-determined number of turns is exceeded. Examples of the partial-input SU, the full-input SU, and the baseline SU are shown in Figure 2. The natural language generation part of the SU uses different templates for simulating the above types of SU, and for making the SU provide direct vs. conversational responses.

In order to make a decision, the SU only has access to the information that a real user would have, i.e., status of devices, previous user action and intention, and previous system action. Likewise the system policy has only access to the information that a real system would have, i.e., status of devices, previous system action and intention, and previous user action. The system policy and the SU do not have access to each other's internal representations and can deal with misunderstandings and missing information by using techniques such as reference resolution. Thus both the system policy and the SU are very realistic and can work with other users and system policies respectively.

Partial-input SU
System configuration: no misunderstandings, conversational, explicit
(Washer in bathroom: status is off)
(Washer in kitchen: status is off)
(TV in bedroom: status is on)
U: System, can you please turn on the washer in 20 min?
S: In which room?
U: Bathroom.
S: Sure thing. The washer in the bathroom will be turned on in 20 min.
(Washer in bathroom: status scheduled to become on in 20 min)
U: Okay. Turn off the TV in the bedroom.
S: Okey dokey. I've turned off the TV in the bedroom.
(TV in bedroom: status is off)
U: Thanks.
S: You're very welcome.

Full-input SU
System configuration: misunderstandings, direct, implicit
(Washer in bathroom: status is off)
U: Can you please turn on the washer in the bathroom in 50 min?
S: Your request has been taken care of.
(Washer in bathroom: status scheduled to become on in 15 min)
U: No. I said turn on the washer in the bathroom in 50 min.
S: Apologies. Done.
(Washer in bathroom: status scheduled to become on in 50 minutes)

Baseline SU
System configuration: misunderstandings, conversational, implicit
(TV in bedroom: status is off)
(Speaker in bedroom: status is off)
U: Turn on the TV in the bedroom. Turn on the speaker in the bedroom.
S: I'm on it. Done.
(TV in bedroom: status is off)
(Speaker in bedroom: status is on)
U: Turn on the TV in the bedroom. Turn on the speaker in the bedroom.
S: I'm on it. Done. The speaker in the bedroom is already on.
(TV in bedroom: status is on)

Figure 2: Example simulated dialogues using a partial-input SU, a full-input SU, and a baseline SU.

The next step was to evaluate our SUs. We collected ratings from MTurk with regard to user rationality, user naturalness, and perceived task completion for both real users (WOz dialogues) and SUs. User rationality, user naturalness, and perceived task completion (whether the MTurkers thought that the tasks were accomplished) roughly evaluate the logic, natural language component, and persistence

User Type	Ration -ality	Natural -ness	Perceived Task Completion
All SUs	5.72	5.54	97.22%
Full-input SU	5.81	5.69	97.63%
Partial-input SU	5.75	5.63	97.35%
Baseline SU	5.60	5.29	96.70%
Real user	5.54	5.39	92.23%

Table 4: Comparison between real user behavior and simulated user behavior based on MTurkers' ratings.

of the SU. For this evaluation we used 24 real dialogues (3 for each one of the 8 possible system configurations) and 72 simulated dialogues (3 for each one of the 8 possible system configurations and 3 SU types). We recruited 100 MTurkers and each MTurker had to rate 15 dialogues with regard to the rationality and naturalness of the user on a 7-point Likert scale (1:low, 7:high), and perceived task completion. Note that we also asked the MTurkers to evaluate the system in terms of intelligence, naturalness, personality, friendliness, their enjoyment, overall quality, and whether they would recommend the system to others. Again, the MTurkers had no information about the features that the SU and the system policy were using to generate dialogues (misunderstandings vs. no misunderstandings, conversational vs. direct system/user behavior, implicit vs. explicit system behavior). They were just instructed to rate the dialogues based on their own interpretation of the rated aspects, and they did not know that these were simulated dialogues.

The results with regard to user rationality, user naturalness, and perceived task completion are shown in Table 4. The difference in terms of naturalness between the full-input and the baseline model is significant (p=0.0003). Also, the difference in terms of naturalness between the partial-input and the baseline model is significant (p=0.0054). Finally the full-input SU is significantly different from the real user in terms of rationality (p=0.0412) and naturalness (p=0.0099). The rest of the differences are not significant. For testing statistical significance, we used the two-tailed unpaired t-test with Holm-Bonferroni correction for repeated comparisons.

We also calculated pairwise Pearson correlations between all features in our simulated data and MTurkers' ratings. Table 5 shows the correlations of the MTurkers' ratings with the conversational and misunderstandings binary features in the simulated dialogues. Correlations with the explicit feature were not significant.

5. Dialogue Quality Evaluation Functions

We performed regression experiments to come up with dialogue quality evaluation functions that are predictive of human ratings. We randomly split our corpus of 186 WOz dialogues into a training set and a test set (138 dialogues for training and 48 for testing) making sure that dialogues from the same real user did not appear in both the training and the test sets. We applied linear regression to the training set, calculated our evaluation functions, and then

MTurkers' Rating	Feature	Pearson's r
Intelligence	Misund	-0.92***
Naturalness	Convers	0.46***
Naturalness	Misund	-0.58***
Overall quality	Misund	-0.91***
Recommend	Misund	-0.88***
Personality	Convers	0.72***
Enjoyment	Convers	0.48***
Enjoyment	Misund	-0.60***
Friendliness	Convers	0.67***
User rationality	Misund	-0.65***
User naturalness	Convers	0.32**
User naturalness	Misund	-0.52***
Perceived task completion	Misund	-0.46***

Table 5: Pearson's r correlations of MTurkers' ratings with dialogue features, simulated dialogues (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, with Holm-Bonferroni correction).

Rating	Function	RMSE
Intellig	0.1*Conv+0.2*Expl-2.2*Mis+6.4	1.32
Natural	0.8*Conv-0.1*Expl-0.9*Mis+5.8	1.09
Over qual	0.4*Conv+0.3*Expl-2.0*Mis+5.8	1.43
Recomm	0.4*Conv+0.2*Expl-2.1*Mis+5.7	1.94
Personal	1.3*Conv-0.4*Expl-0.5*Mis+5.2	1.19
Enjoym	0.6*Conv+0.1*Expl-1.7*Mis+5.6	1.63
Friendl	0.9*Conv-0.1*Expl-0.6*Mis+5.7	1.10

Table 6: Evaluation functions trained on WOz dialogues and real users' ratings and tested on WOz dialogues and real users' ratings (Conv: conversational, Expl: explicit, Mis: misunderstandings binary features).

measured how these evaluation functions performed on the test set (i.e., how predictive they were of the ratings in the test set). To do that we calculated the root mean square error (RMSE) as shown in Equation (1) where n is the number of dialogues, $Rat_{iPredicted}$ is the predicted "Rating" for dialogue i (calculated by our evaluation function), and $Rat_{iActual}$ is the actual "Rating" for dialogue i. Obviously the lower the RMSE the better. The RMSE scale is from 0 to 6 because the ratings were on a scale from 1 to 7.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Rat_{iPredicted} - Rat_{iActual})^2} \quad (1)$$

Similarly, we randomly split our corpus of 72 simulated dialogues into a training and a test set (48 dialogues for training and 24 for testing). Then we applied linear regression to the training set, calculated our evaluation functions, and then measured how these evaluation functions performed on the test set using the RMSE.

Table 6 shows the evaluation functions trained on the training set of WOz dialogues and real users' ratings, as a function of the conversational, explicit, and misunderstandings

Rating	Function	RMSE
Intellig	0.2*Conv+0.3*Expl-1.0*Mis+4.9	0.90
Natural	1.1*Conv-0.1*Expl-0.1*Mis+3.8	0.97
Over qual	0.5*Conv+0.2*Expl-0.9*Mis+4.8	0.72
Recomm	0.5*Conv+0.3*Expl-1.0*Mis+4.6	0.90
Personal	1.0*Conv+0.1*Expl-0.3*Mis+4.1	0.97
Enjoym	0.6*Conv+0.2*Expl-0.7*Mis+4.2	0.89
Friendl	0.6*Conv-0.1*Expl-0.4*Mis+4.6	0.80

Table 7: Evaluation functions trained on WOz dialogues and MTurkers' ratings and tested on WOz dialogues and MTurkers' ratings (Conv: conversational, Expl: explicit, Mis: misunderstandings binary features).

Rating	Function	RMSE
Intellig	0.2*Conv+0.0*Expl-1.9*Mis+5.8	0.35
Natural	0.5*Conv+0.1*Expl-0.6*Mis+5.2	0.28
Over qual	0.3*Conv+0.0*Expl-2.0*Mis+5.8	0.34
Recomm	0.4*Conv-0.0*Expl-1.8*Mis+5.2	0.34
Personal	1.2*Conv+0.2*Expl-0.1*Mis+3.7	0.55
Enjoym	0.7*Conv+0.1*Expl-0.8*Mis+4.3	0.37
Friendl	1.0*Conv+0.2*Expl-0.3*Mis+4.5	0.55

Table 8: Evaluation functions trained on simulated dialogues and MTurkers' ratings and tested on simulated dialogues and MTurkers' ratings (Conv: conversational, Expl: explicit, Mis: misunderstandings binary features).

features. It also shows the RMSE values when these evaluation functions are applied on the test set of WOz dialogues and real users' ratings. Table 7 shows the evaluation functions trained on the training set of WOz dialogues and MTurkers' ratings, as a function of the conversational, explicit, and misunderstandings features. It also shows the RMSE values when these evaluation functions are applied on the test set of WOz dialogues and MTurkers' ratings. Table 8 shows the evaluation functions trained on the training set of simulated dialogues and MTurkers' ratings, as a function of the conversational, explicit, and misunderstandings features. It also shows the RMSE values when these evaluation functions are applied on the test set of simulated dialogues and MTurkers' ratings.

Table 9 shows the results in terms of RMSE values when we apply on the test set of WOz data and real users' ratings the evaluation functions that were trained on simulated data and MTurkers' ratings from Table 8 (SIM column), the evaluation functions that were trained on WOz data and MTurkers' ratings from Table 7 (OBS column), and the evaluation functions that were trained on WOz data and real users' ratings from Table 6 (REAL column). Thus the fourth column of Table 9 is the same as the RMSE column of Table 6.

For testing statistical significance, we used the two-tailed paired t-test with Holm-Bonferroni correction for repeated comparisons. It is interesting that the RMSE for intelligence-SIM is lower than the RMSE for intelligence-OBS ($p < 0.05$), and is not significantly different from the RMSE for intelligence-REAL. Also, the RMSE for

Rating	SIM	OBS	REAL
Intelligence	1.35	1.56	1.32
Naturalness	1.23	1.86	1.09
Overall quality	1.45	1.64	1.43
Recommend	1.98	2.06	1.94
Personality	1.61	1.50	1.19
Enjoyment	1.77	1.77	1.63
Friendliness	1.45	1.66	1.10

Table 9: RMSE values for different evaluation functions when all are tested on the WOz data and real users' ratings. SIM: means trained on simulated data and MTurkers' ratings, OBS: means trained on WOz data and MTurkers' ratings, and REAL: means trained on WOz data and real users' ratings.

Comparison	Statistical Signific
Intelligence-SIM vs. Intelligence-OBS	$p < 0.05$
Intelligence-SIM vs. Intelligence-REAL	NS
Intelligence-OBS vs. Intelligence-REAL	$p < 0.05$
Naturalness-SIM vs. Naturalness-OBS	$p < 0.001$
Naturalness-SIM vs. Naturalness-REAL	NS
Naturalness-OBS vs. Naturalness-REAL	$p < 0.001$
Overall-SIM vs. Overall-OBS	NS
Overall-SIM vs. Overall-REAL	NS
Overall-OBS vs. Overall-REAL	NS
Recommend-SIM vs. Recommend-OBS	NS
Recommend-SIM vs. Recommend-REAL	NS
Recommend-OBS vs. Recommend-REAL	NS
Personality-SIM vs. Personality-OBS	NS
Personality-SIM vs. Personality-REAL	NS
Personality-OBS vs. Personality-REAL	NS
Enjoyment-SIM vs. Enjoyment-OBS	NS
Enjoyment-SIM vs. Enjoyment-REAL	NS
Enjoyment-OBS vs. Enjoyment-REAL	NS
Friendliness-SIM vs. Friendliness-OBS	NS
Friendliness-SIM vs. Friendliness-REAL	$p < 0.05$
Friendliness-OBS vs. Friendliness-REAL	$p < 0.05$

Table 10: Comparisons (regarding statistical significance) between RMSE values for different evaluation functions when all are tested on the WOz data and real users' ratings; two-tailed paired t-test with Holm-Bonferroni correction. SIM: means trained on simulated data and MTurkers' ratings, OBS: means trained on WOz data and MTurkers' ratings, REAL: means trained on WOz data and real users' ratings, and NS: means no significant difference.

intelligence-OBS is significantly different from the RMSE for intelligence-REAL ($p < 0.05$). The RMSE for naturalness-SIM is lower than the RMSE for naturalness-OBS ($p < 0.001$), and is not significantly different from the RMSE for naturalness-REAL. The RMSE for naturalness-OBS is significantly different from the RMSE for naturalness-REAL ($p < 0.001$). For intelligence and naturalness the predicted ratings that we get by training on simulated data and MTurkers' ratings are quite close to the

predicted ratings we get by training on WOz data and real users' ratings, which is very encouraging. For recommend the system to others and enjoyment none of the models does well, which is not surprising given that these are hard to capture aspects of the interaction. In terms of personality, the OBS evaluation function is doing better than the SIM evaluation function and they are both worse than the REAL evaluation function. With regard to friendliness the SIM model performs worse than the REAL model ($p < 0.05$), and the OBS model also performs worse than the REAL model ($p < 0.05$). Finally, with regard to overall quality the performance of the SIM model is quite close to the performance of the REAL model, but overall there are no significant differences among the RMSE values of all models. Table 10 summarizes the statistical significance of all pairwise comparisons.

6. Conclusion

We collected WOz dialogues in the IoT domain. We asked the users participating in these dialogues to rate the system regarding intelligence, naturalness, personality, friendliness, their enjoyment, overall quality, and whether they would recommend the system to others. Then we asked MTurkers to rate these dialogues on the same aspects. We also generated simulated dialogues between dialogue policies and simulated users and asked MTurkers to rate them again on the same aspects. We developed evaluation functions based on features from the simulated dialogues and the MTurkers' ratings, the WOz dialogues and the MTurkers' ratings, and the WOz dialogues and the WOz participants' ratings. We applied all these evaluation functions to a held-out portion of our WOz dialogues. For intelligence and naturalness the evaluation functions trained on simulated data and MTurkers' ratings performed significantly better than the evaluation functions trained on WOz data and MTurkers' ratings, and similarly to the evaluation functions trained on WOz data and real users' ratings. Also, for overall quality the evaluation functions trained on simulated data and MTurkers' ratings performed similarly to the evaluation functions trained on WOz data and real users' ratings. This is encouraging because it shows that at least for these three conversational aspects just training evaluation functions on simulated data could be sufficient. Recommend the system and enjoyment were hard to predict for all evaluation functions. For personality and friendliness the evaluation functions trained on the WOz data and the real users' ratings did much better than the other evaluation functions.

7. Acknowledgments

This work was funded in part by Samsung Electronics Co., Ltd., and partly supported by the U.S. Army. Statements and opinions expressed do not necessarily reflect the policy of the United States Government, and no official endorsement should be inferred.

8. Bibliographical References

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies – why and how. *Knowledge-Based Systems*, 6(4):258–266.

El Asri, L., He, J., and Suleman, K. (2016). A sequence-to-sequence model for user simulation in spoken dialogue systems. In *Proc. of Interspeech*, pages 1151–1155, San Francisco, California, USA.

Foster, M. E., Giuliani, M., and Knoll, A. (2009). Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proc. of ACL*, pages 879–887, Suntec, Singapore.

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. (2015). DeltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proc. of ACL (Short Papers)*, pages 445–450, Beijing, China.

Georgila, K., Henderson, J., and Lemon, O. (2005). Learning user simulations for information state update dialogue systems. In *Proc. of Interspeech*, pages 893–896, Lisbon, Portugal.

Georgila, K., Henderson, J., and Lemon, O. (2006). User simulation for spoken dialogue systems: Learning and evaluation. In *Proc. of Interspeech*, pages 1065–1068, Pittsburgh, Pennsylvania, USA.

Georgila, K., Gordon, C., Choi, H., Boberg, J., Jeon, H., and Traum, D. (2018). Toward low-cost automated evaluation metrics for Internet of Things dialogues. In *Proc. of IWSDS*, Singapore.

Gordon, C., Georgila, K., Choi, H., Boberg, J., and Traum, D. (2018). Evaluating subjective feedback for Internet of Things dialogues. In *Proc. of SemDial:AixDial*, pages 64–72, Aix-en-Provence, France.

Gordon, C., Yanov, V., Traum, D., and Georgila, K. (2019). A Wizard of Oz data collection framework for Internet of Things dialogues. In *Proc. of SemDial:LondonLogue*, pages 168–170, London, UK.

Guo, F., Metallinou, A., Khatri, C., Raju, A., Venkatesh, A., and Ram, A. (2017). Topic-based evaluation for conversational bots. In *Proc. of NIPS Workshop on Conversational AI: Today's Practice and Tomorrow's Potential*, Long Beach, California, USA.

Hastie, H. (2012). Metrics and evaluation of spoken dialogue systems. In Oliver Lemon et al., editors, *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, pages 131–150. Springer.

Henderson, J., Lemon, O., and Georgila, K. (2008). Hybrid reinforcement/supervised learning of dialogue policies from fixed datasets. *Computational Linguistics*, 34(4):487–511.

Hone, K. S. and Graham, R. (2000). Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Journal of Natural Language Engineering*, 6(3-4):287–303.

Jeon, H., Oh, H. R., Hwang, I., and Kim, J. (2016). An intelligent dialogue agent for the IoT home. In *Proc. of the AAAI Workshop on Artificial Intelligence Applied to Assistive Technologies and Smart Environments*, pages 35–40, Phoenix, Arizona, USA.

Jurčíček, F., Thomson, B., and Young, S. (2012). Reinforcement learning for parameter estimation in statistical spoken dialogue systems. *Computer Speech and Language*, 26(3):168–192.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proc. of EMNLP*, pages 2122–2132, Austin, Texas, USA.

Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). On the evaluation of dialogue systems with next utterance classification. In *Proc. of SIGDIAL*, pages 264–269, Los Angeles, California, USA.

Misu, T., Georgila, K., Leuski, A., and Traum, D. (2012). Reinforcement learning of question-answering dialogue policies for virtual museum guides. In *Proc. of SIGDIAL*, pages 84–93, Seoul, South Korea.

Paksima, T., Georgila, K., and Moore, J. D. (2009). Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. In *Proc. of SIGDIAL*, pages 1–10, London, UK.

Pietquin, O. and Hastie, H. (2013). A survey on metrics for the evaluation of user simulations. *Knowledge Engineering Review*, 28(1):59–73.

Robinson, S., Roque, A., and Traum, D. (2010). Dialogues in context: An objective user-oriented evaluation approach for virtual human dialogue. In *Proc. of LREC*, pages 64–71, Valletta, Malta.

Schatzmann, J. and Young, S. (2009). The hidden agenda user simulation model. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):733–747.

Schatzmann, J., Georgila, K., and Young, S. (2005). Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proc. of SIGDIAL*, pages 45–54, Lisbon, Portugal.

Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review*, 21(2):97–126.

Traum, D. R., Robinson, S., and Stephan, J. (2004). Evaluation of multi-party virtual reality dialogue interaction. In *Proc. of LREC*, pages 1699–1702, Lisbon, Portugal.

Walker, M., Kamm, C., and Litman, D. (2000). Towards developing general models of usability with PARADISE. *Journal of Natural Language Engineering*, 6(3-4):363–377.