What Went Wrong and Why? Diagnosing Situated Interaction Failures in the Wild

Sean Andrist, Dan Bohus, Ece Kamar, Eric Horvitz

Microsoft Research
Redmond, WA 98052
{sandrist, dbohus, eckamar, horvitz}@microsoft.com

Abstract. Effective situated interaction hinges on the well-coordinated operation of a set of competencies, including computer vision, speech recognition, and natural language, as well as higher-level inferences about turn taking and engagement. Systems often rely on a set of hand-coded and machine-learned components organized into several sensing and decision-making pipelines. Given their complexity and inter-dependencies, developing and debugging such systems can be challenging. "In-the-wild" deployments outside of controlled lab conditions bring further challenges due to unanticipated phenomena, including unexpected interactions such as playful engagements. We present a methodology for assessing performance, identifying problems, and diagnosing the root causes and influences of different types of failures on the overall performance of a situated interaction system functioning in the wild. We apply the methodology to a dataset of interactions collected with a robot deployed in a public space inside an office building. The analyses identify and characterize multiple types of failures, their causes, and their relationship to overall performance. We employ models that predict overall interaction quality from various combinations of failures. Finally, we discuss lessons learned with such a diagnostic methodology for improving situated systems deployed in the wild.

Keywords: situated interaction, human-robot interaction, dialog systems, integrative AI, failure diagnosis.

1 Introduction

Systems designed to engage in physically situated language interaction with human users in the open world, such as social robots and conversational agents, rely on multiple competencies to interact effectively. These systems combine speech recognition and vision pipelines with models for higher-level inferences (e.g., about presence, intentions and attention, speakers and addressees, etc.) and interaction planning.

An important set of challenges in the development of these systems arises from the need to coordinate multiple heterogeneous components for sensing, reasoning, and acting. The interdependencies among these components can lead to development, refinement, and maintenance problems [1]. Many components are machine learned and thus have nondeterministic behaviors. Often, components are tightly coupled in multistage processing pipelines, making it difficult to diagnose errors and assign blame. For example, when failures occur in a speech processing pipeline, it is difficult to pinpoint exactly which of the relevant sub-components (e.g., echo cancellation, voice-activity detection, speech recognition, language understanding, etc.) is to blame and where efforts for improving the system should be focused. Since components may be trained on the erroneous outputs of other components, making improvements to individual components can lead to novel failures downstream and degraded performance [2].

Another set of challenges in these systems arises from the nature and diversity of interactions that may occur in the real world [3, 4]. When designing and testing situated systems in a laboratory environment, subjects are typically instructed to carry out a specific task with the system. Functionality and failures are assessed in relation to a battery of subjective and objective task-specific metrics. However, once deployed in the wild, the specific task goals assumed in the design and study of the system may not align with the goals of people approaching or engaging with the system. Users may attempt actions that are out-of-domain and therefore difficult to handle. Even when a user is interacting within the system's intended domain, there can be wide variability in interaction styles and attitudes. Especially during first encounters with the robotic system, many users may be driven by curiosity about the system, rather than a real task-centric need; they may playfully test the system's capabilities. The types of failures occurring in an interactive system may change based on the nature of the interaction.

We present an annotation-based methodology combining observer and system-expert views to investigate failures in a deployed situated interactive system. The methodology is composed of a workflow of annotations and analyses for identifying the type of interaction users engage in, assessing the overall interaction quality, identifying specific problem types on a turn-by-turn basis, and diagnosing the root causes of these problems along with their impact on the system's overall performance. Root causes are associated with specific components identified by the system expert. We investigate relationships between observed failures, their component causes, and the overall interaction quality with the goal of diagnosing the most important system problems.

We showcase the methodology through a case study with a directions-giving humanoid robot. The methodology allows us to gain insight into the system's behavior and performance towards guiding future engineering efforts. The analyses show that problems with the content of robot utterances explain overall interaction quality better than problems with timing. Across different components, speech recognition and engagement component failures are most informative in understanding overall quality, and data-driven models that consider all component failures are capable of predicting quality scores close to the level of a human annotator. The analyses also show that the type of the interaction that happens in the wild is an important factor in error diagnosis affecting not only interaction quality but also the types and origins of problems.

2 Background

Prior relevant research on failure diagnosis has proposed the use of crowdworkers to understand bottlenecks and to perform blame assignment in single-shot machine

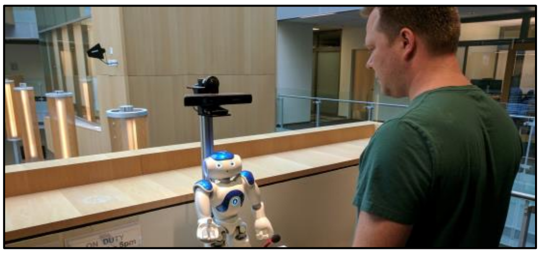


Figure 1. *The Directions Robot interacting with a user*

learning pipelines [2, 5]. However, this approach hinges on the ability to rapidly simulate system execution to test potential fixes in the loop, and is thus not viable for temporally extended situated interactions seen with human-robot interaction.

A large body of work has addressed the topics of evaluation and blame assignment in spoken dialog systems [6, 7]. Schmitt et al. investigate *interaction quality* at arbitrary points in the dialogue by utilizing external observer annotations [11]. The PARADISE framework utilizes multiple regression analysis to model the relationship between objective task success, dialogue costs, and user satisfaction of a spoken dialogue system [12]. Methods for detecting errors directly from recognition results, inferring previous errors from user reactions to the system, and predicting possible future errors based on dialogue turns have also been explored [13]. Walker et al. [14] assessed the performance of error detection models using multiple components of a dialogue system, finding the best results when all components were considered together (ASR, NLP, Dialogue Manager, etc.). Our work complements and extends these efforts by considering larger heterogeneous robot systems with multimodal interaction competencies and by reasoning about system behavior in the wild with respect to different interaction types.

3 Directions Robot

The experimental basis for the work reported in this paper is the Directions Robot [8] (Figure 1), which couples a Nao robot with off-board sensors and processing. The system uses an external wide-angle RGB camera and a Kinect sensor in speech and vision processing pipelines which allow it to reason about conversational engagement [9] (i.e., determine who is engaged in an interaction) and turn-taking (i.e., determine when it should speak) in multiparty interaction. The robot can understand requests for directions to people's offices by name or number, as well as to common areas (e.g., kitchen, bathrooms, etc.). When speaking, the robot coordinates speech with head movements and arm gestures. The system is deployed outside the elevators on the third floor in our building. Traffic in this area includes both building residents and visitors. We placed signs near the robot and inside the elevators briefly describing the robot capabilities, as well as the policies for data collection and opting out.

Interactions that naturally occur in this space can be classified into several distinct types. A subset of interactions is characterized by users having a real need to seek directions, which we refer to as **In-Domain-Real** interactions. However, previous research with this system has revealed that a large portion of interactions are not based in an actual need for directions [10]. Some interactions arise out of curiosity and desire to test the systems with directions requests (**In-Domain-Exploring**). Others diverge entirely from seeking directions, e.g., users asking the robot about the weather (**Out-of-Domain**). Moreover, due to various failures in face detection and engagement modeling, the robot sometimes falsely starts interactions with "imagined" people or with people who did not wish to engage. We refer to these as **Falsely-Initiated** interactions. Different user intentions and interaction types have influences on the behavior of the system and on considerations for a successful interaction.

4 Annotating Failures

We collected a dataset of 173 interactions over a period of eleven days with the Directions Robot. The average interaction duration was 25 seconds, with the longest interaction lasting 164 seconds. 141 (82%) of the interactions involved a single engaged participant, while 32 (18%) involved two or more engaged participants.

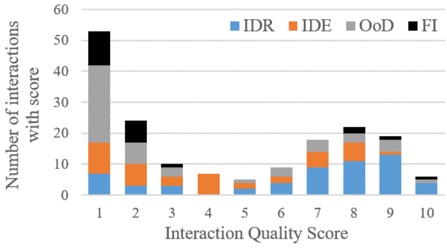
Our methodology involves annotating the dataset across four different dimensions: *interaction type, overall interaction quality, problem types* (detailed content and timing problems), and *component causes* (underlying failure points in the system causing the problems). The first three annotations (interaction type, interaction quality, and problem types) reflect externally observable aspects and perceptions of the robot's behavior. These annotations were performed by a trained annotator with a background in linguistics and were based on observing the recorded interactions from the robot's perspective via custom log visualization tools. To check for inter-annotator agreement, a second annotator labeled a subset containing 20% (35) of the interactions in the dataset.

In contrast, component causes capture the internal functioning of the system. These annotations require expert technical knowledge of the system and were conducted by the first and second author. Each expert labeled 60% of the problem type occurrences, with a 20% overlap allowing for computing agreement.

4.1 Interaction Types

Each interaction was labeled as one of the four types described above: In-Domain-Real, In-Domain-Exploring, Out-of-Domain, and Falsely-Initiated. The annotation scheme used the following as a guiding question: "was this at any point an in-domain conversational interaction in which an actor seemed to genuinely need directions in the building?" Inter-annotator agreement on the 20% subset of data was high; Cohen's $\kappa = .80$.

We found that the robot falsely initiated 13% (23) of the 173 interactions. The rest were roughly evenly split among Out-of-Domain (29%), In-Domain-Exploring (25%) and In-Domain-Real (32%). The percentage of users with real intentions is higher than was found in previous work (19% [10]) due to our conscious decision to bias the annotation scheme toward labeling the interaction as In-Domain-Real given *any* evidence of a real intention at any point in the interaction. We believe this scheme aligns with the proper strategy for deployed robots, namely to assume real intentions unless highly confident otherwise. Figure 2 (right) shows the basic statistics of the dataset, divided



	# Interactions	Avg Length (s)	Avg Score
All	173 (100%)	25	4.5
IDR	56 (32%)	26	6.5
IDE	43 (25%)	42	4.1
OoD	51 (29%)	18	3.3
FI	23 (13%)	9	2.7

Figure 2. ***Left***: *Histogram of interaction quality scores broken down by interaction type: In-Domain-Real (IDR), In-Domain-Exploring (IDE), Out-of-Domain (OoD), and False-Interaction (FI).* ***Right***: *Table of interaction counts and durations.*

also by interaction types. Falsely initiated interactions stand apart in terms of duration, with a much shorter average length of 9 seconds.

4.2 Interaction Quality

The Interaction Quality annotation captures the overall performance of the robot in the interaction. Annotators viewed each interaction in its entirety and rated the robot's performance on a scale from 1 ("completely broken") to 10 ("perfect"). The annotator was instructed to take a holistic view as an external-observer view of the interaction and not focus solely on the robot's or participants' perspectives.

The average score across all interactions is 4.5. Assigning semantic meaning to absolute scores between the two endpoints of the scale is difficult, but we can make relative comparisons of average scores. As Figure 2 shows, the distribution of scores is bimodal, with 50% of interactions assigned less than or equal to 3 and 38% of interactions scoring greater than or equal to 7. Decomposition by interaction type (Figure 2) shows that scores vary across interaction types. Falsely initiated interactions receive the lowest average score (2.7), and in-domain interactions driven by a real need have the highest average score (6.5). We observe lower interaction quality when the nature of the interaction does not agree with the system goals, highlighting challenges with the system running in the wild. We assessed inter-annotator agreement by computing Spearman's ρ correlation coefficient for the ordinal scale (ρ=0.78).

4.3 Problem Types

The third annotation aims to provide an in-depth look at the problems that arise on a turn-by-turn basis during the interactions. Success in situated interactions hinges on producing the right responses with the right timing. We therefore annotate both content and timing problems. A content problem occurs when the robot's utterance is not appropriate given the current context, e.g., the robot misunderstands the user and provides directions to the incorrect destination. A timing problem occurs when the timing of the robot's utterance is incorrect, e.g., a robot starts speaking at a point when the user had not yet released the floor, overlapping with the user.

We segmented each interaction into units, each starting from the beginning of a system dialog act and extending until the start of the next. This segmentation resulted in

995 total units. To assess each unit, annotators were asked: (1) "Was this the right time for the robot to speak?" and (2) "Was the content of this utterance appropriate?" Annotators were instructed to imagine the “gold standard” of what a competent human might do if they were stepping into the robot’s place at that moment in the interaction. They were encouraged to consider the context of everything that had occurred prior, including previous errors and misunderstandings. Overall, 439 units (44%) were found to have a content problem and 277 units (28%) were found to have a timing problem. 154 units (15%) were labeled as problematic in both content and timing.

The annotation scheme for content problems contained thirteen sub-labels (a subset shown in Figure 4a). Frequent content problems include cases when the robot: asks the user to repeat themselves (*AskRepeat*); asks what the user needs after they had already specified that (*AskWhat*); asks the user for confirmations due to low speech recognition confidence (*Confirm*). The content of greetings and farewells can also be problematic when the robot attempts to engage users who do not wish to interact (*Initiate-FalseEngagement*) or when it attempts to interact with users who no longer wish to engage (*NoUserEngagement*, *MissingFarewell*).

The annotation scheme for timing problems had eleven sub-labels (subset shown in Figure 4b). These failures are often related to turn taking, such as pausing too long after the user stops speaking (*TooLongPause*). Conversely, the robot sometimes steals the conversational floor and talks over the user (*StealFloorSpeaking*). Timing can also be broken at the beginning and ending of engagements. False engagements may be triggered before any user wishes to engage (*InitiateFalseEngagement*) and disengaging farewells may come too late because the user has already left (*LateFarewell*).

We observed high annotator agreement on the labeling of individual content problems (Cohen’s κ = 0.64), whereas agreement was slightly lower for the assessment of timing problems (Cohen’s κ = 0.59). Figure 4a and 4b shows the frequency of top content and timing problems broken down by interaction type. Falsely-Initiated interactions have the largest proportions of content (57%) and timing (57%) problems, while In-Domain-Real interactions exhibit the smallest (35% content, 24% timing). The type of the interaction affects the types of frequent content and timing problems. Individually, content and timing issues related to engagement appear to be most prevalent in the falsely initiated interactions, and content failures around confirmation appear more in In-Domain-Exploring interactions (15%) than In-Domain Real interactions (9%).

4.4 Component Causes

After all units were annotated for content and/or timing problems, the first and second authors annotated each problematic unit for the most likely cause of the indicated problem. This assessment required expert knowledge of the architecture and inner workings of the system. The system architecture (Figure 3) was abstracted into several pipelines, each comprising a number of components such as the face tracker, engagement model, dialog model, echo cancellation, voice activity detection, speech recognition, turn-taking model, etc. In any given turn, multiple system components may fail simultaneously. For feasibility of labeling and modeling, the expert annotator identified the earliest possible component in a pipeline that failed.

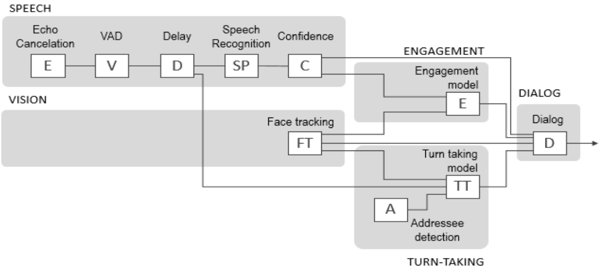


Figure 3. *Pipelines (gray boxes) of components (white boxes) comprising the system.*

Occasionally the robot disengages early without providing a farewell utterance, resulting in having no utterance unit to annotate. Because this is an important error to capture, the annotators were instructed to provide an additional indicator for the overall timing appropriateness of the robot's disengagement from each interaction: *TooEarly*, *TooLate*, *Good*, or *Other*. This assessment was highly subjective and ambiguous, resulting in low overall agreement (Cohen's κ = 0.19). When we collapse the labels into *TooEarly* (31 interactions) and *NotTooEarly* (142 interactions), Cohen's κ increases to 0.52. This indicator is included in the list of component causes.

Figure 4c depicts the most common component causes, aggregated across both content and timing problems, and split by interaction type. Component failures in the engagement model, speech recognition, and the dialog model emerged as the most selected causes overall. The analyses show that the components that cause the most problems vary across interaction types. For example, failures in the engagement model were noted most often for falsely initiated interactions (44%). Speech recognition errors were much more prevalent in the other interaction types, particularly in the In-Domain-Exploring (12%) and Out-of-Domain (12%) interactions. This analysis of understanding how different components lead to problems for different interaction types can guide decisions about which components to invest in to improve performance. For example, if there is a higher cost of making mistakes for falsely initiated interactions, the best strategy would be to improve the engagement component.

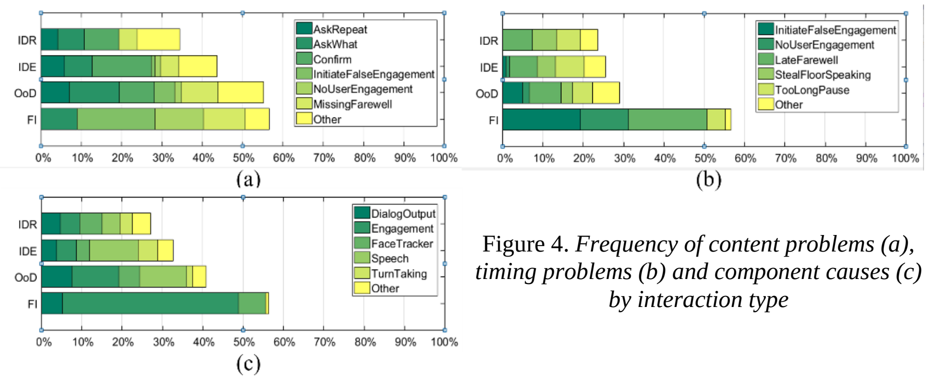


Figure 4. *Frequency of content problems (a), timing problems (b) and component causes (c) by interaction type*

Table 1. *Mean squared error (MSE) from logistic regression models trained on various problem types and component failures, across interaction types.*

	All Interactions	IDR	IDE	OoD	FI
Baseline	10.48	8.74	7.23	9.49	8.76
%Content	5.72	5.39	4.77	4.58	6.64
%Timing	8.64	8.04	7.48	6.89	7.53
Speech	8.83	6.25	5.17	8.46	8.76
Engagement	9.03	8.11	6.93	9.04	7.58
Speech + Engagement	6.66	6.52	5.09	7.14	7.58
Face Tracking	10.48	8.63	6.33	8.86	8.76
Addressee Detection	10.30	7.83	7.31	9.49	8.76
All Components + Interaction Type	4.65	5.26	6.28	3.98	7.34

5 Explaining Interaction Quality

We turn our attention next to understanding how types of problems and component failures affect the interaction score. We construct logistic regression models that predict the ordinal interaction quality score (scaled down to the 0-1 interval) from various problems indicators. We compute mean squared error (MSE) in a leave-one-out cross-validation process and compare with a baseline model that simply predicts the mean rating.

We begin by looking at how well the ratio of content and timing problems explain the overall quality score. The results, shown in Table 1, indicate that percentage of content problems in an interaction achieves a MSE of 5.72. The percentage of timing problems is less informative (MSE=8.64).

To determine which component failures are most informative for predicting the interaction score, we construct single variable models based on the percentage of failures from each component in the pipeline. The most informative component failures are Speech (MSE=8.83) and Engagement (MSE=9.03). Interestingly, when we add these two features together, we achieve a much better MSE of 6.66. A model that leverages all component causes, plus an indicator for the interaction type, achieves a mean squared error of 4.65. We can compare this model to how well the secondary annotator's scores predict the primary annotator's scores. On those 35 annotations, the secondary annotator achieves a prediction accuracy MSE of 4.14, while our model achieves MSE=3.67. Although our model performs well here, it is important to note that the model is trained on the primary annotator's scores (using the leave-one-out method) and that the rating task itself is subjective among human annotators.

We also train and test the same logistic regression models separately on each interaction type (Table 1). This analysis yields several interesting insights, e.g., that failures in the engagement component are particularly informative for falsely initiated interactions, failures in addressee detection are important for InDomain-Real interactions, and failures in face tracking are important when users are exploring the system.

Ideally, robots should be able to assess interaction quality during the interaction, rather than only at the very end, allowing it to take steps to improve the interaction or acknowledge that things are going wrong. To assess how this might be possible, we

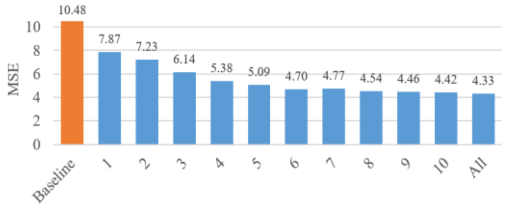


Figure 5. *Increasing prediction accuracy of the AllComponents+InteractionType Model from the first N units of the interaction. Baseline simply predicts the mean.*

applied our best performing model (AllComponents+InteractionType) to the first N units of interactions (Figure 5). Predictive power increases as more units are considered, but reasonable accuracy seems to be achievable approximately after five units.

6 Conclusion and Future Work

We presented a methodology and study on error diagnosis in situated interactive systems. The methodology combines observer and system-expert annotations to collect signals about problems occurring in the interaction. We employ quantitative analyses and predictive modelling to study relationships among interaction types, problems, their causes, and overall interaction quality. Overall, we found that content problems have a stronger impact on overall performance than timing problems. The major identified causes of problems are failures in the speech recognition and engagement components. Models that consider all component failures, along with a feature indicating the type of interaction users are engaging in (real, exploring, out-of-domain, or false), reach a performance approaching that of another human annotator in predicting interaction quality.

The annotation-based methodology and study has several limitations that highlight future directions for research. First, the approach requires a large set of detailed annotations from both observers and system-experts. Our experience indicates that these annotation efforts could be further streamlined. Future work may investigate automating some of these annotations via machine learning. Since different interaction types exhibit different profiles of failures and causes, another future direction is developing automated methods for inferring the interaction type at runtime [10].

When identifying component causes, we have focused on the earliest component in the pipeline that fails. As highlighted in earlier research, problems in integrative systems may be caused simultaneously by multiple components [2]. Due to entanglement issues and non-monotonic error propagation, fixing a component early in a pipeline does not necessarily lead to improved system performance. Furthermore, from the user's perspective, the critical failure point might not align with the component identified by our methodology. Future work is needed to understand how the methodology can be applied iteratively to address these issues.

The annotations of interaction quality and problem types rely on a third-person view rather than a first-person view, which would require direct user input and might differ. For example, an exploring user might be satisfied with the interaction even in the presence of content failures but an external observer might rate the interaction poorly. Future work can aim to understand how these views differ and design interaction methods for capturing the first-person assessment without overburdening users.

The dataset we have collected provides an initial basis for deeper analyses and exploration in blame assignment. More sophisticated analyses may leverage additional interaction context, include temporal aspects, and lead to a richer understanding about which types of failures occur together or which types of failures are indicators for future problems. These understandings promise to be useful for enabling systems to perform prediction of forthcoming failures and to engage in self-repair and recovery.

References

1. D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J. Crespo, and D. Dennison. (2015). “Hidden technical debt in machine learning systems”, in Advances in Neural Information Processing Systems (pp. 2503-2511).
2. B. Nushi, E. Kamar, D. Kossmann, and E. Horvitz. (2017). “On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems”, in Proceedings of AAAI 2017.
3. N. Mirnig, A. Weiss, G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, B. Granström, and M. Tscheligi. (2013). “Face-to-face with a robot: What do we actually talk about?”, In International Journal of Humanoid Robotics, 10(01).
4. G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, and B. Granström. (2012). “Furhat at robotville: A robot head harvesting the thoughts of the public through multi-party dialogue”, In Proceedings of the International Conference on Intelligent Virtual Agents.
5. D. Parikh and C. L. Zitnick. (2011). “Human-Debugging of Machines”, in The Second Workshop on Computational Social Science and the Wisdom of Crowds (NIPS 2011).
6. S. Georgiladakis, G. Athanasopoulou, R. Meena, J. Lopes, A. Chorianopoulou, E. Palogiannidi, E. Iosif, G. Skantze, A. Potamianos. (2016). “Root Cause Analysis of Miscommunication Hotspots in Spoken Dialogue Systems”, In Interspeech 2016 (pp. 1156-1160).
7. M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. (1997). “PARADISE: A framework for evaluating spoken dialogue agents”, In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 271-280).
8. D. Bohus, C.W. Saw, and E. Horvitz. (2014). “Directions Robot: In-the-Wild Experiences and Lessons Learned”, in Proceedings of AAMAS’2014, Paris, France.
9. D. Bohus, and E. Horvitz. (2014). “Managing Human-Robot Engagement with Forecasts and … um … Hesitations”, in Proceedings of ICMI’2014, Istanbul, Turkey.
10. S. Andrist, D. Bohus, Z. Yu, and E. Horvitz. (2016). “Are You Messing with Me?: Querying about the Sincerity of Interactions in the Open World”, in Proceedings of HRI 2016. IEEE Press, Piscataway, NJ, USA, 409-410.
11. Schmitt, A., Schatz, B., & Minker, W. (2011). “Modeling and predicting quality in spoken human-computer interaction”. In *Proceedings of the SIGDIAL 2011 Conference* (pp. 173-184).
12. Walker, Marilyn A., Litman, D. J., Kamm, C. A., & Abella, A. "PARADISE: A framework for evaluating spoken dialogue agents." *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997.
13. Meena, Raveesh, Lopes, J., Skantze, G., & Gustafson, J. "Automatic Detection of Miscommunication in Spoken Dialogue Systems." In *Proc. of SIGDIAL*. 2015.
14. Walker, Marilyn, Langkilde, I., Wright, J., Gorin, A., & Litman, D. "Learning to predict problematic situations in a spoken dialogue system: experiments with how may I help you?." In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. 2000.