# Toward Low-Cost Automated Evaluation Metrics for Internet of Things Dialogues

Kallirroi Georgila, Carla Gordon, Hyungtak Choi,
Jill Boberg, Heesik Jeon, David Traum

USC Institute for Creative Technologies

SAMSUNG

# Internet of Things (IoT)

- The Internet of Things (IoT) is a 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
  - all together via an integrated app
  - by a smart assistant via human-system dialogue interaction

# Samsung's Vision: Internet of Things video

- https://www.youtube.com/watch?v=9rYqkERzY00&app=desktop

# Motivation

- Evaluation of dialogue systems is expensive, requiring high-cost subjective human judgements
- Existing automated metrics do not specifically address challenges in IoT dialogues
- Our goal is to develop automated evaluation functions that can predict human ratings in the IoT domain

# Outline

- Dialogue evaluation
- Challenges in the IoT domain
- Dialogue corpus
- Annotations
- Collection of human ratings via crowdsourcing
- Dialogue quality evaluation functions
- Conclusion and future work

# Outline

- **Dialogue evaluation**
- Challenges in the IoT domain
- Dialogue corpus
- Annotations
- Collection of human ratings via crowdsourcing
- Dialogue quality evaluation functions
- Conclusion and future work

# Dialogue evaluation methods

- Evaluation with human users
- Evaluation with simulated users
- Automated evaluation functions that are predictive of subjective human judgements
  - PARADISE framework (Walker et al., 2000) – optimize a desired quality (e.g., user satisfaction) by formulating it as a linear combination of a variety of metrics (e.g., task success and dialogue length)

# Dialogue evaluation metrics

- Subjective human judgements, e.g., user satisfaction, perceived task completion, etc.
- Quantitative automated metrics, e.g., task completion, dialogue length, word error rate, etc.
  - Quantitative automated similarity metrics that require a reference corpus, e.g., BLEU, METEOR, ROUGE, precision, recall, accuracy

# Problems with similarity metrics

- Context matters in dialogue
  - A different response early on will result in an entirely different dialogue from the dialogue in the reference corpus and make comparisons meaningless
- Do not address dialogue coherence
- Penalize rather than reward novelty and creativity
- Do not correlate well with human judgements of dialogue quality (Liu et al., 2016)

# Examples of problem with BLEU score

Original dialogue

U: Turn on the light in the bedroom.
S: Did you say the bathroom?
U: No, the bedroom.
S: Turning on the light in the bedroom.
U: Thanks.

Example #1

Low BLEU score but better than original!

U: Turn on the light in the bedroom.
S: Turning on the light in the bedroom.
U: You are the best!

Example #2

High BLEU score but incoherent

U: Turn on the light in the bedroom.
S: Did you say the bathroom?
U: Yes, the bedroom.
S: Turning on the light in the livingroom.
U: Thanks.

# Outline

- Dialogue evaluation
- **Challenges in the IoT domain**
- Dialogue corpus
- Annotations
- Collection of human ratings via crowdsourcing
- Dialogue quality evaluation functions
- Conclusion and future work

# Challenges in the IoT domain

U: Play music in 10 minutes.

S: In which room shall I play music?

U: Bedroom please.

S: As you wish. I will play music in the bedroom in 10 minutes.

U: Thanks. Please turn on the air-conditioner there too.

S: Okay, the air-conditioner in the bedroom is now on.

U: Wait, I want the air-conditioner on in 10 minutes, not now.

Timing-related misunderstanding

U: Turn on the TV in the bedroom.

S: Done. Anything else?

U: Yes, turn on the light.

S: I've turned on the light in the bedroom.

U: No, turn on the light here in the kitchen.

S: I've turned on the light in the kitchen. Anything else?

U: Nope. That's all for now.

Location-related misunderstanding and potentially undesired side effects (2 lights on instead of 1)

# Challenges in the IoT domain (continued)

U: Set the temperature to 30 degrees.

S: I can't do that while the cycle is running.

U: Okay. Set the program to wool.

S: I can't do that either. The cycle will have to be stopped first.

U: Okay. Stop the cycle. Then set the temperature to 30 degrees, change the program to wool, and start the cycle again.

S: Sure thing.

U: Move the TV up one channel.

S: Moving the TV up one channel.

U: No, I mean down one channel.

S: Got it. I'm moving the TV down one channel.

The context of device state and actions performed is important

# Outline

- Dialogue evaluation
- Challenges in the IoT domain
- **Dialogue corpus**
- Annotations
- Collection of human ratings via crowdsourcing
- Dialogue quality evaluation functions
- Conclusion and future work

# Our dialogue corpus

- 6200 dialogues between a smart home assistant and a user
  - Written by linguists and designed to capture phenomena encountered in real human-system dialogues
    - Example issues: speech recognition errors, misunderstandings, clarification requests, timing, context and scheduling issues, etc.
  - Different levels of task complexity
    - One device at a time (washer, speaker, bulb, TV, air-conditioner)
    - Multiple devices at the same time (e.g., the air-conditioner and the TV).
    - Multiple devices of the same type (e.g., TV in the bedroom, TV in the kitchen, and TV in the guest room)
- 232 annotated dialogues used in our experiments

# Statistics of corpus used in experiments (232 dialogues)

| Dialogue feature | Mean | Standard deviation |
|---|---|---|
| Number of tasks per dialogue | 1.41 | 0.68 |
| Number of system turns per dialogue | 2.80 | 0.98 |
| Number of user turns per dialogue | 2.80 | 0.98 |
| Number of all turns per dialogue | 5.60 | 1.96 |
| Number of system words per dialogue | 14.08 | 8.01 |
| Number of user words per dialogue | 15.32 | 6.00 |
| Number of all turns per dialogue | 29.40 | 12.15 |
| Average number of system words per utterance | 5.14 | 2.61 |
| Average number of user words per utterance | 5.69 | 1.88 |
| Average number of all words per utterance | 5.41 | 1.81 |

# Outline

- Dialogue evaluation
- Challenges in the IoT domain
- Dialogue corpus
- **Annotations**
- Collection of human ratings via crowdsourcing
- Dialogue quality evaluation functions
- Conclusion and future work

# Automated annotations

- Number of system and user turns per dialogue
- Number of total words from system and user per dialogue
- Average number of words per system and user utterance in a dialogue
- Number of occurrences of specific words and expressions e.g., “yes/yeah/yep/yup”, “no/nope”, “ok/okay”, “alright/all right”, “done”, “system”, “thanks/thank you”, “good/great”, “not at all”, “sorry/apologize/apologies”, etc.

# Manual annotations of system utterances

| Big Category | Middle Category | Small Category | Annotation Type | Annotation Type description | Example Interaction | |
|---|---|---|---|---|---|---|
| | | | | | User | System |
| SYSTEM Utterance | Assess Action | Action | A-something | System does something | "Please connect the speaker" | "I am connecting the speaker" |
| | | | A-nothing | System does nothing | "Please connect the speaker" | "Which speaker do you want me to connect?" |
| | | | A-valid | System does requested thing | "Please turn on the kitchen light." | "I am turning on the kitchen light" |
| | | | A-invalid | System does not do requested thing | "Please turn on the kitchen light." | "I am turning on the porch light" |
| | Response type | Describe Current Understanding | CU-confirm | confirm request before doing | "Please turn on the kitchen light." | "Do you want me to turn the kitchen light on?" |
| | | | CU-restate | restate understanding | "Please turn on the kitchen light." | "Do you want me to turn the kitchen light on?" |
| | | | CU-lack | describe lack of understanding | "Please turn it on" | "Sorry I don't know what you want." |
| | | Action Acknowledgement | AA-past | Action specified, past | "Please turn on the kitchen light." | "The kitchen light has been turned on." |
| | | | AA-present | Action specified, present | "Please turn on the kitchen light." | "I am turning on the kitchen light." |
| | | | AA-future | Action specified, future | "Please turn on the kitchen light in 5 minutes." | "I will turn on the kitchen light in 5 minutes." |
| | | | ANS | Action not specified | "Please turn on the kitchen light." | "Done." |
| | | | AI | Action impossible | "Open the washer door." | "Sorry, I can't open it before finishing the cycle" |
| | | Specify State | SS-done-E | explicit, action done | "Please turn on the kitchen light." | "The kitchen light is now on." |
| | | | SS-done-I | implicit, action done | "Please turn on the kitchen light." | "It is now on." |
| | | | SS-NA-E | explicit, not applicable | "Please turn on the kitchen light." | "The kitchen light is already on." |
| | | | SS-NA-I | implicit, not applicable | "Please turn on the kitchen light." | "It is already on." |

# Manual annotations of user utterances

| Big Category | Middle Category | Small Category | Annotation Type | Annotation Type description | Example Interaction | |
|---|---|---|---|---|---|---|
| | | | | | System | User |
| USER Utterance | Request | Request Action | RA-dev | specified device | | "Turn off the washer." |
| | | | RA-loc | specified location | | "Turn on the rinsing for the wash machine in guest room." |
| | | | RA-time | specified time | | "Start the net connection for washing machine in the dining room in 5 minutes." |
| | | | RA-temp | specified temp | | "System, decrease temperature to 40" |
| | | | RA-end-state | specified end state | | "I feel like listening to the music you play." |
| | | | RA-other | specified other | | "Connect the speaker to bluetooth." |
| | | | RA-action | only action specified | | |
| | Response | Response to System | RS-yes | yes | "Are you sure? The washing is still in progress." | "Yes." |
| | | | RS-no | no | "Do you want me to open the door of the washer in the kitchen?" | "No." |
| | | | RS-null | null | | |
| | | | RS-restate | restate request | "I'm not sure if I understood properly. Which TV should I link to the network?" | "telly which is in hall." |
| | | | RS-correct | correct value | "Garage AC will be connected to WIFI in 50 minutes." | "In 15 minutes." |
| | | | RS-decline | decline further action | "Anything else?" | "No, thanks." |
| | | | RS-param | provide parameter(s) | "When should I switch it on?" | "Today at 3 pm." |
| | Pleasantries | Pleasantries | P-greet | greeting | | "Hello dear system!" |
| | | | P-thank | thank you | | "Thank you." |

# Manual annotation examples

| | SYSTEM | | | | | | | | | USER | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Assess Actio | Response Type | | | | | Specificit | Linguistic features | | Utterance type | | | Specificity |
| | Action type | Describe current understanding | Acknowledge action | Specify State | Request | Other | level of specificit | Register | Grammaticality | Request action | Response to System | Pleasantries | level of specificity |
| Washer1 | | | | | | | | | | | | | |
| Let's initiate the rinse programme on the washer. | | | | | | | | | | RA-action RA-dev | | | explicit |
| No problem. Rinsing mode selected. | A-something A-valid | | AA-present | | | | explicit | Reg-conv | gram | | | | |
| Please start washing in 10 minutes. | | | | | | | | | | RA-action RA-time | | | explicit |
| I will start washing in 10 minutes. | A-nothing | | AA-future | | | | explicit | Reg-direct | gram | | | | |
| Washer2 | | | | | | | | | | | | | |
| Run the rising cycle please. | | | | | | | | | | RA-action | | | |
| Sorry I don't know what you want. | A-nothing | CU-lack | | | | | N/A | Reg-conv | gram | | | | |
| Set rinse cycle. | | | | | | | | | | RA-action | | | |
| I'm afraid we can not set the rinse mode during | A-nothing | | AA-AI | | | | explicit | Reg-conv | gram | | | | |
| Stop the washing now. | | | | | | | | | | RA-action RA-time | | | |
| Ok. it's stopped. | A-something A-valid | | | SS-unclear | | | implicit | Reg-direct | gram | | | | |
| Now set the rinse cycle. | | | | | | | | | | RA-action RA-time | | | |
| Rinse mode set. | A-something A-valid | | | SS-unclear | | | implicit | Reg-direct | gram | | | | |
| Washer3 | | | | | | | | | | | | | |
| Switch washing mode to rinse for bathroom washer. | | | | | | | | | | RA-action RA-dev RA-loc | | | explicit |
| I've set rinsing mode for bathroom washing machine. | A-something A-valid | | AA-past | | | | explicit | Reg-direct | gram | | | | |
| Decrease temperature by 15. | | | | | | | | | | RA-action RA-temp | | | explicit |

# Additional manual or semi-automated annotations

- Number and list of devices in the dialogue and their locations
- Number and list of misunderstandings in the dialogue
- Number and list of immediate and scheduling tasks in the dialogue

U: System, decrease the washer temperature to 40.
S: Changing temperature to 50.
U: No, change it to 40.
S: Of course. Setting temperature to 40.
U: Start the speaker in the bedroom in 10 minutes.
S: The bedroom speaker is scheduled to start in 10 minutes.

Devices considered: 2 [Dev=washer, DevLoc=null, Dev=speaker, DevLoc=bedroom]
Immediate tasks: 1 [Dev=washer, DevLoc=null, SetTemp=40]
Scheduling tasks: 1 [Dev=speaker, DevLoc=bedroom, Start, 10min]
Misunderstandings: 1 [Dev=washer, DevLoc=null, SetTempMis=50, SetTemp=40]

# Outline

- Dialogue evaluation
- Challenges in the IoT domain
- Dialogue corpus
- Annotations
- **Collection of human ratings via crowdsourcing**
- Dialogue quality evaluation functions
- Conclusion and future work

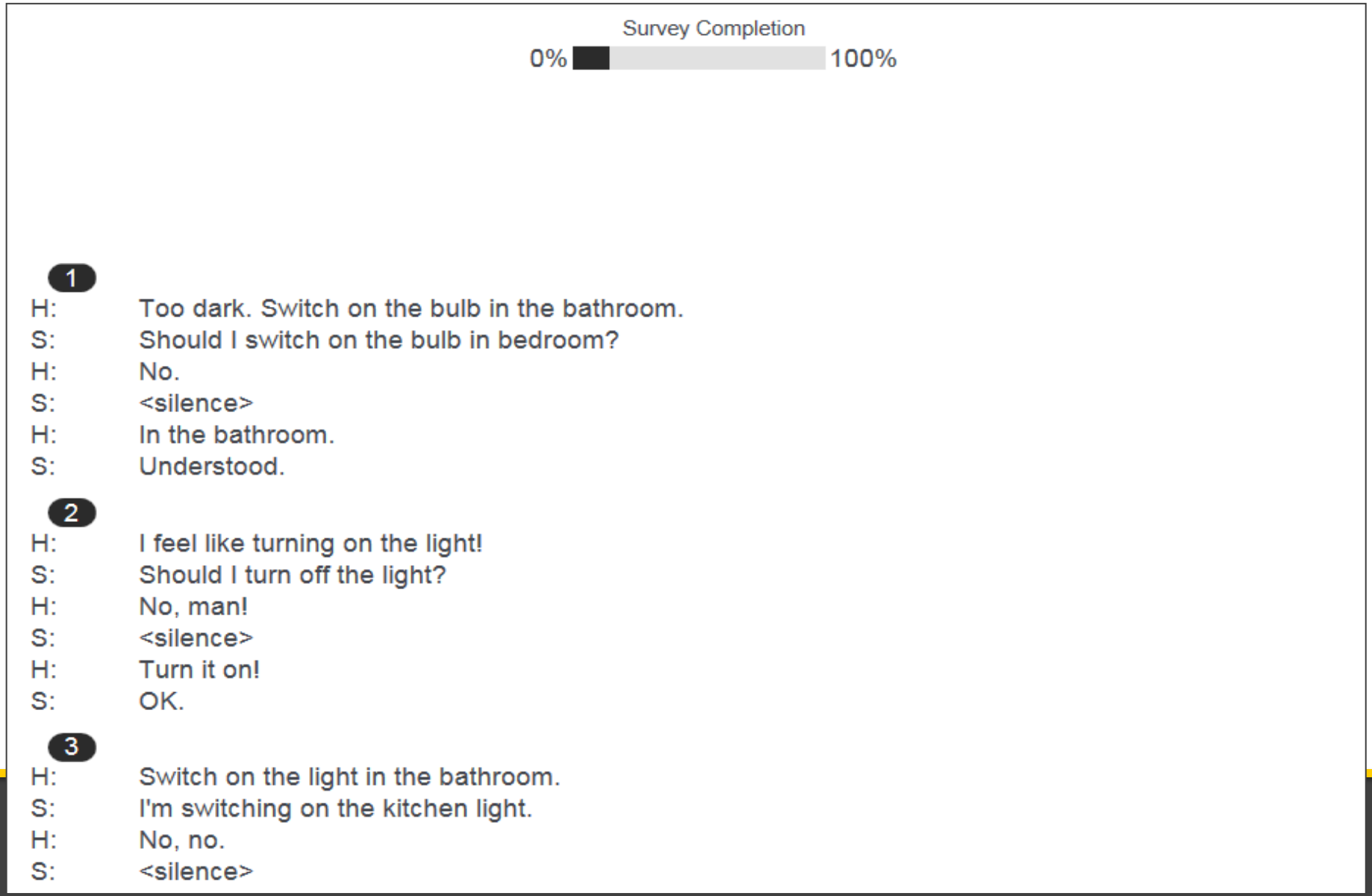
# Collection of human ratings via crowdsourcing

- 199 raters from Amazon Mechanical Turk
- 4 tasks
- For each task raters were presented with 8 sets of 5 dialogues and asked to rank them (best dialogue to worst dialogue) based on which system they would most like to interact with

# Crowdsourced tasks description

- **Task1:** All dialogues in a set had the same task goal and device (turn on TV, connect speaker to Bluetooth, etc.)
- **Task 2:** Dialogues in a set represented a mix of task goals and devices
- **Task 3:** Dialogues in a set represented a mix of task goals and devices, and raters were presented with a description of the pre-conversation system status (state and location of each device before the dialogue starts)
- **Task 4:** Dialogues in a set represented a mix of task goals and devices, and raters were presented with a description of the pre-conversation and post-conversation system status (state and location of each device before the dialogue starts and after the dialogue ends)

# Design of crowdsourced experiment (survey on Qualtrics)



# Outline

- Dialogue evaluation
- Challenges in the IoT domain
- Dialogue corpus
- Annotations
- Collection of human ratings via crowdsourcing
- **Dialogue quality evaluation functions**
- Conclusion and future work

## Generate dialogue scores from rankings

- Compare each dialogue $D_x$ with each other $D_y$
- For each pairs of dialogues ($D_x$, $D_y$) in all rankings
  - If $D_x$ ranked higher than $D_y$
    - then $Win_x$ ++, $Lose_y$ ++
    - Else $Lose_x$ ++, $Win_y$ ++
- Score $D_x = Win_x / (Win_x + Lose_x)$

# Selected Pearson correlations of "Score" with features (***: p<0.001, **: p<0.01, *: p<0.05)

| Dialogue feature | Pearson's r |
|---|---|
| **Number of misunderstandings** | **-0.76*** ** |
| **Misunderstandings exist or not? (Binary)** | **-0.77*** ** |
| **Number of system confirmation requests** | **-0.50*** ** |
| **System confirmation requests exist or not? (Binary)** | **-0.50*** ** |
| Number of system requests for more information | 0.27*** |
| System requests for more information exist or not? (Binary) | 0.28*** |
| **Number of silence occurrences** | **-0.67*** ** |
| **Silence occurrences exist or not? (Binary)** | **-0.68*** ** |
| **Number of times the system does nothing (A-nothing)** | **-0.54*** ** |
| System does nothing (A-nothing) or not? (Binary) | -0.36*** |
| Number of times the system does something invalid (A-invalid) | -0.33*** |
| System does something invalid (A-invalid) or not? (Binary) | -0.33*** |
| System has a conversational style or not? (Binary) | 0.17** |

# Selected Pearson correlations of misunderstandings with features (***: p<0.001, **: p<0.01, *: p<0.05)

| Dialogue feature | Pearson'r r (for counts) | Pearson's r (for binary values) |
|---|---|---|
| Occurrences of silence | 0.42*** | 0.44*** |
| Occurrences of "I mean/I meant" | 0.22*** | 0.24*** |
| Occurrences of "I said" | 0.25*** | 0.29*** |
| Occurrences of "no/nope" | 0.66*** | 0.64*** |
| Number of system turns per dialogue | 0.58*** | - |
| Number of user turns per dialogue | 0.58*** | - |
| Number of system words per dialogue | 0.55*** | - |
| Number of user words per dialogue | 0.47*** | - |
| System does nothing (A-nothing) | 0.47*** | 0.33*** |
| System does something invalid (A-invalid) | 0.41*** | 0.40*** |
| User specifies the location of a device (RA-location) | 0.17** | 0.24*** |

# Regression experiments

- We experimented with variations of the following features (highly correlated with “Score”):
  - Number of misunderstandings (**Misund**)
  - Misunderstandings exist or not? (Binary) (**MisundBin**)
  - Number of system confirmation requests (**Confirm**)
  - System confirmation requests exist or not? (Binary) (**ConfirmBin**)
  - Number of system requests for more information (**Info**)
  - System requests for more information exist or not? (Binary) (**InfoBin**)
  - System has a conversational style (at least half of system responses are annotated as “Reg-conv”) or not? (Binary) (**ConvBin**)

# Regression experiments (continued)

- Split the corpus in training (75%) and test (25%) sets
- Apply linear regression to the training set and derive the evaluation functions
- Measure how the evaluation functions perform on the test set (how predictive they are of actual human ratings) by calculating the root mean square error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(PredictedScore(i) - ActualScore(i))^2}$$

n: number of dialogues

PredictedScore(i): score for dialogue i calculated by evaluation function

ActualScore(i): score for dialogue i derived from human ratings

# Best evaluation functions in terms of RMSE

| Description | Evaluation function | RMSE |
|---|---|---|
| Reward-based function | 100*Task_success-5*Num_system_turns | 0.5224 |
| Misund | -0.21*Misund+0.62 | 0.0902 |
| MisundBin | -0.23*MisundBin+0.62 | 0.0920 |
| Misund+Confirm | -0.18*Misund-0.05*Confirm+0.62 | 0.0919 |
| MisundBin+ConfirmBin | -0.20*MisundBin-0.05*ConfirmBin+0.63 | 0.0929 |
| **Misund+Info** | **-0.20*Misund+0.02*Info+0.61** | **0.0899** |
| MisundBin+InfoBin | -0.23*MisundBin+0.01*InfoBin+0.62 | 0.0919 |
| Misund+Info+ConvBin | -0.21*Misund+0.02*Info-0.01*ConvBin+0.62 | 0.0911 |
| MisundBin+InfoBin+ConvBin | -0.23*MisundBin+0.01*InfoBin-0.01*ConvBin+0.62 | 0.0924 |
| Misund+Confirm+Info | -0.18*Misund-0.05*Confirm+0.01*Info+0.62 | 0.0915 |
| MisundBin+ConfirmBin+InfoBin | -0.20*MisundBin-0.05*ConfirmBin+0.01*InfoBin+0.63 | 0.0928 |

# Outline

- Dialogue evaluation
- Challenges in the IoT domain
- Dialogue corpus
- Annotations
- Collection of human ratings via crowdsourcing
- Dialogue quality evaluation functions
- **Conclusion and future work**

# Conclusion

- The evaluation functions that include “misunderstandings”, and to a lesser extent “system confirmation requests”, “system requests for more information”, and “conversational style” are all good predictors of the real “Scores” (derived from the human ratings)
- The reward-based evaluation function resulted in much higher RMSE

# Future work

- Build models that can recreate rankings within a set of dialogues and see whether these derived rankings agree with the actual human rankings in our dataset
- Collect more realistic dialogues in a Wizard of Oz setting
  - Include role of nonlinguistic context (states and actions)
- Develop evaluation functions that are tailored to specific users or groups of users
  - Some raters liked the ability of the system to explicitly state the action that it was about to perform (grounding) and request more information (as a clarification request)
  - Other raters appreciated brevity and preferred more implicit system responses

# Thank you!

- Questions?