User satisfaction in dialogue systems

Hsien-chin Lin

2021.05.14

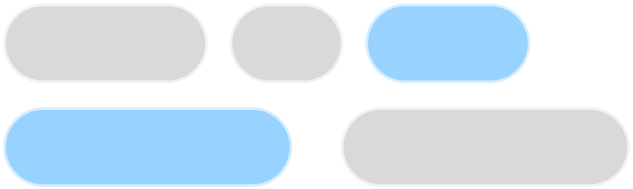
Outline

Heinrich Heine
Universität Düsseldorf

- What is user satisfaction
- How to model user satisfaction
- How to use user satisfaction to improve dialogue systems
- The challenge of using user satisfaction as a reward function

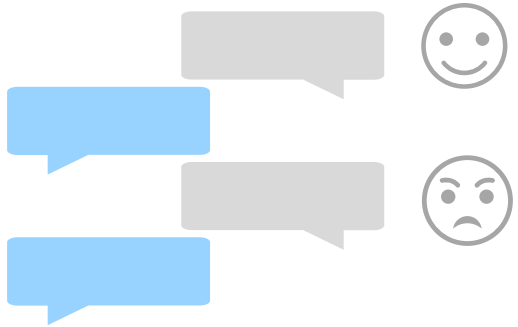
Before we start...

Heinrich Heine
Universität Düsseldorf



Key word extraction

Image classification

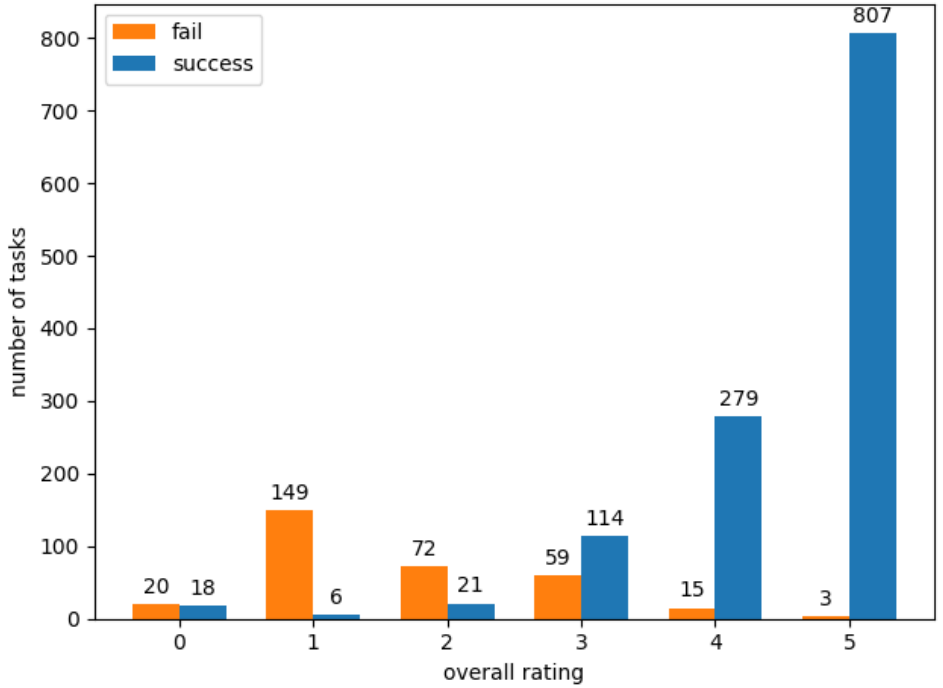


Dialogue system

User-centric dialogue system development

Heinrich Heine
Universität Düsseldorf

- Systems need to consider the users' behaviour
- Users act differently depending on their satisfaction level
- Optimise the policy not only on task success but also on user satisfaction



Dialogue is success

U can iget the **address** of the **simming pool** attraction?

S **Address is pool way**, whitehill road, off newmarket road. You are welcome. Is there anything else I can help you with today?

U **do you know of swimming pool?**

S What type of attractions are you interested in? Sure, there are 4 in that area. It is listed as swimmingpool. Did you need any further assistance today?

U swimming pool?

S It's located in the east. There are 4, anything in particular you are looking for? Would you be interested in visiting a swimmingpool? Did you need any further assistance today?

U what is the address?

S They are located at pool way, whitehill road, off newmarket road. Did you need any further assistance today?

Dialogue is success

U can iget the address of the simming pool attraction?

S Address is pool way, whitehill road, off newmarket road. You are welcome. Is there anything else I can help you with today?

U do you know of swimming pool?

S **What type of attractions are you interested in?** Sure, there are 4 in that area. It is listed as swimmingpool. Did you need any further assistance today?

U swimming pool?

S It's located in the east. There are 4, anything in particular you are looking for? Would you be interested in visiting a swimmingpool? Did you need any further assistance today?

U what is the address?

S They are located at pool way, whitehill road, off newmarket road. Did you need any further assistance today?

Dialogue is success but overall rating is 2

U can iget the address of the simming pool attraction?

S Address is pool way, whitehill road, off newmarket road. You are welcome. Is there anything else I can help you with today?

U do you know of swimming pool?

S What type of attractions are you interested in? Sure, there are 4 in that area. It is listed as swimmingpool. **Did you need any further assistance today?**

U swimming pool?

S It's located in the east. There are 4, anything in particular you are looking for? Would you be interested in visiting a swimmingpool? **Did you need any further assistance today?**

U what is the address?

S They are located at pool way, whitehill road, off newmarket road. **Did you need any further assistance today?**

How to model user satisfaction

Heinrich Heine
Universität Düsseldorf

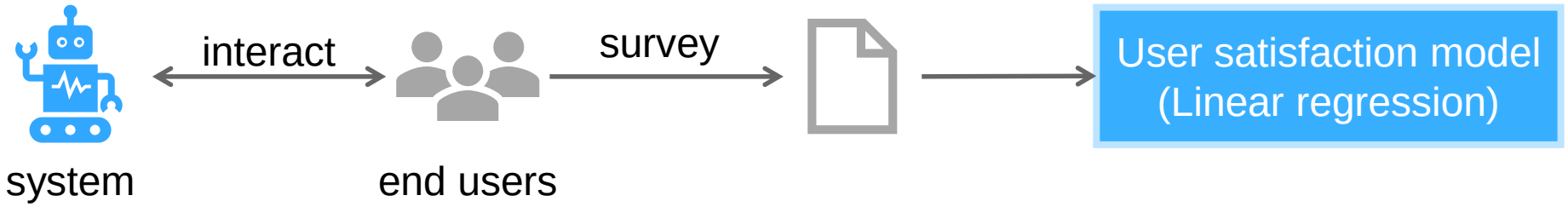
- Dialogue level user satisfaction
  - PARADISE
- Turn level user satisfaction
  - Interaction Quality
  - Response Quality

Dialogue level user satisfaction

Heinrich Heine
Universität Düsseldorf

Model overall rating on the dialogue level

- PARADISE (Walker et al. 1997)
- Task success and dialogue costs contribute to user satisfaction



Problems for dialogue level user satisfaction

Heinrich Heine
Universität Düsseldorf

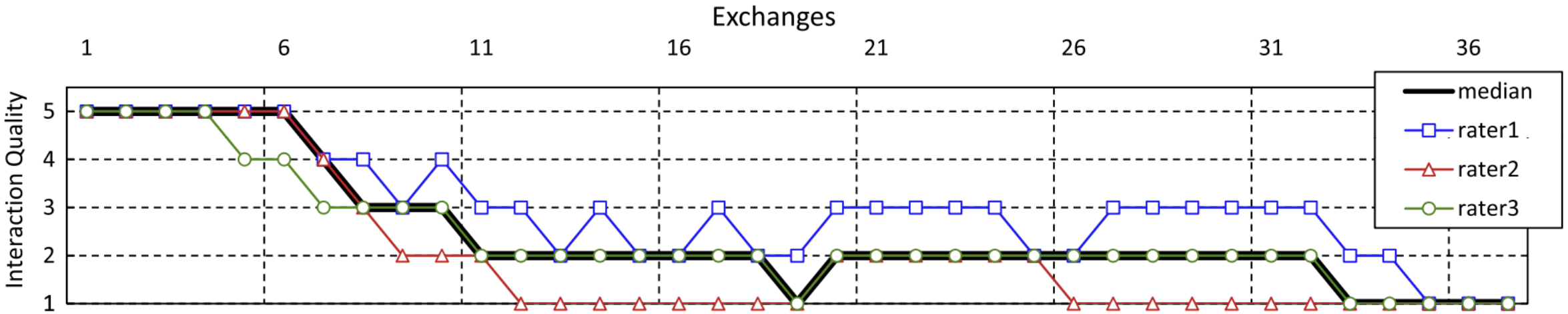
- **Strong variations**: different understanding of functioning interaction
  - Generous v.s picky users
  - Paid users v.s real users
- To label the whole dialogue by expert annotators is difficult
- **Hard to track** real users' satisfaction
- Biased with successful dialogues
  - In commercial systems, the surveys can only be placed for successful dialogues in usual
- Not able to capture the frustration in the intermediate turns

Interaction quality

Heinrich Heine
Universität Düsseldorf

Measure the quality of the interaction up to a certain point in an interaction

- Turn (exchange) level
- A score from 5 to 1



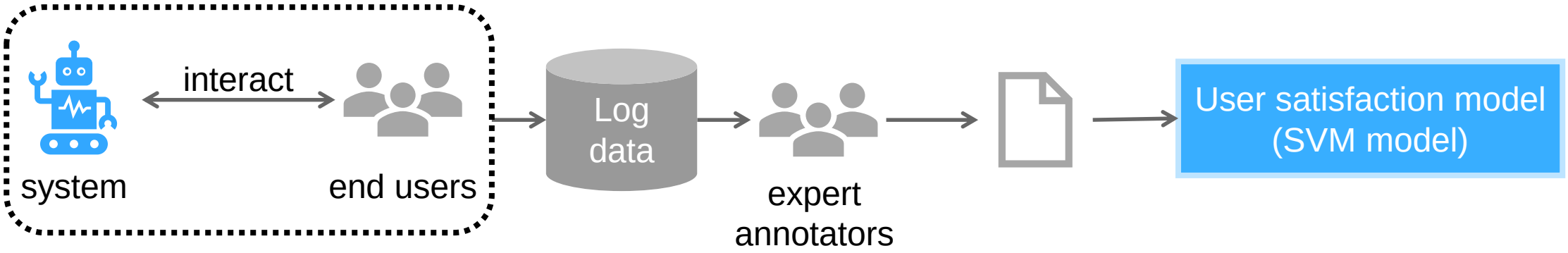
Schmitt, A., & Ultes, S. (2015).

Interaction quality

Heinrich Heine
Universität Düsseldorf

Labeled by experts

- The data labeled by experts is more consistent and objective
- No need for interrupting end users



Interaction quality prediction

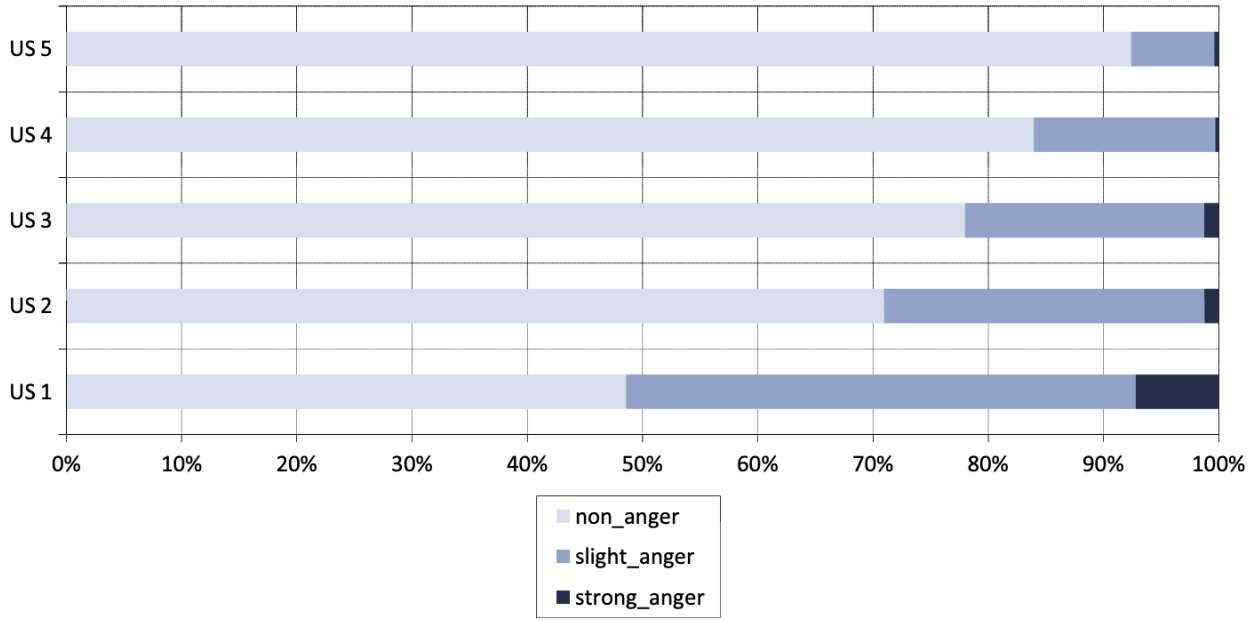
Heinrich Heine
Universität Düsseldorf

- Input features
- Automatic features
  - Automatic speech recognition (ASR): ASR confidence, ...
  - Spoken language understanding (SLU): # of help requests, ...
  - Dialogue manager (DM): loop, ...
- Hand features
  - Dialogue acts
  - Emotion states of the caller
- Log dialogue is from Let's Go bus information system

Parameter	Description
ASRRecognitionStatus	ASR status: *success*, *no match*, *no input*
ASRConfidence	confidence of top ASR results
RePrompt?	is the system question the same as in the previous turn?
ActivityType	general type of system action: *statement*, *question*
Confirmation?	is system action confirm?
MeanASRConfidence	mean ASR confidence if ASR is success
#Exchanges	number of exchanges (turns)
#ASRSuccess	count of ASR status is success
%ASRSuccess	rate of ASR status is success
#ASRRejections	count of ASR status is reject
%ASRRejections	rate of ASR status is reject
{Mean}ASRConfidence	mean ASR confidence if ASR is success
{#}ASRSuccess	count of ASR is success
{#}ASRRejections	count of ASR status is reject
{#}RePrompts	count of times RePromt? is true
{#}SystemQuestions	count of ActivityType is question

Experiment results

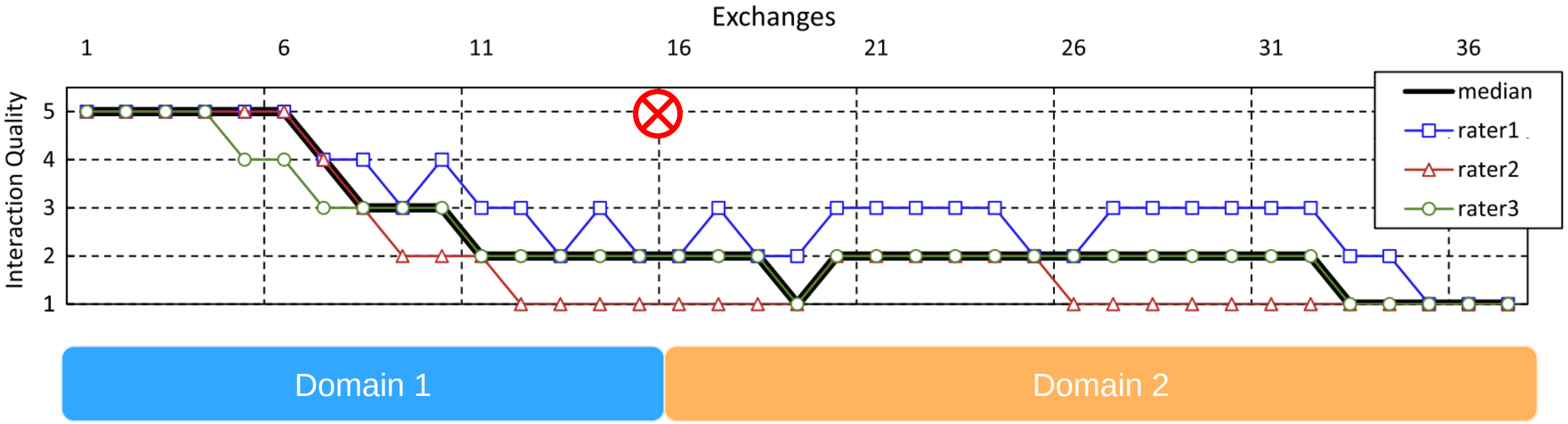
- Interaction quality is correlated with user satisfaction
  - Raters and users experience the dialogue interaction similarly
- User anger
  - The more dissatisfied the users are, the more they express their negative emotion
  - A large proportion of dissatisfied users do not express emotionally



Schmitt, A., & Ultes, S. (2015).

Problems for interaction quality

- It is still necessary to track the dialogue history
- Limited generalizability to multi-domain dialogues



Response quality

Heinrich Heine
Universität Düsseldorf

RQ ratings are provided for each turn independently

- Based on follow-up user feedback
  - frustration expression
  - rephrases user request
- The annotators do not need to deal with full conversation history
- Domain-independent features
  - User request paraphrasing
  - Cohension between response and request
  - ...

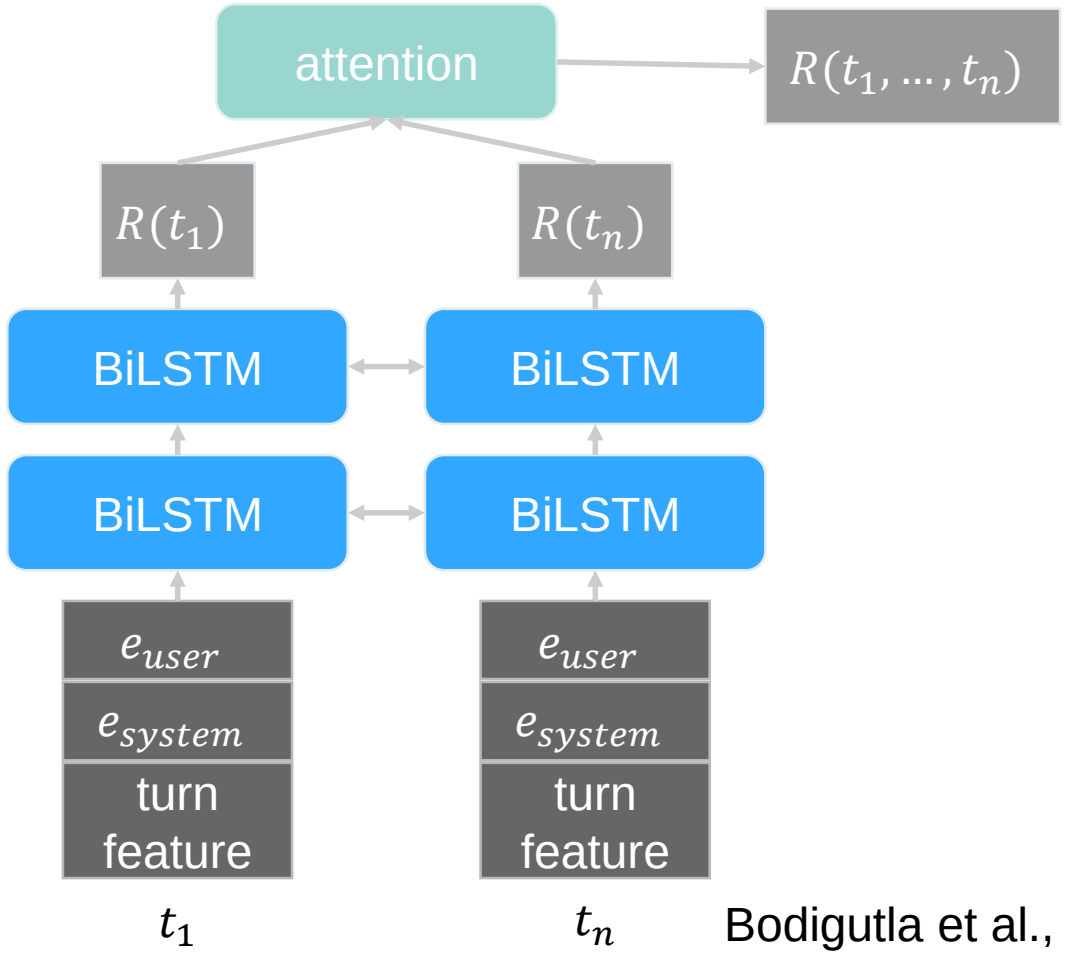
Bodigutla et al., 2019b

Utterance	Turn Rating
User: I would like to book a cab to work	
System: Sure, at what time would you like to be picked up?	**5.0**
User: Eight a.m tomorrow	
System: Could you confirm your cab booking to {office address} at 8 a.m tomorrow?	
You will be charged $50 for this ride	**5.0**
User: No, cancel my ride	
System: Ok, your booking is confirmed.	
Your card ending xxxx has been charged $50	**1.0**

Joint turn and dialogue level user satisfaction

Heinrich Heine
Universität Düsseldorf

- A BiLSTM model optimise turn-level response quality by experts and dialogue level user satisfaction by end users as a multitask learning
- Using a pre-trained sentence encoder to encode user and system utterence
- The model puts more weight on the dissatisfactory turns according to the learnt attention weights



Bodigutla et al., 2020

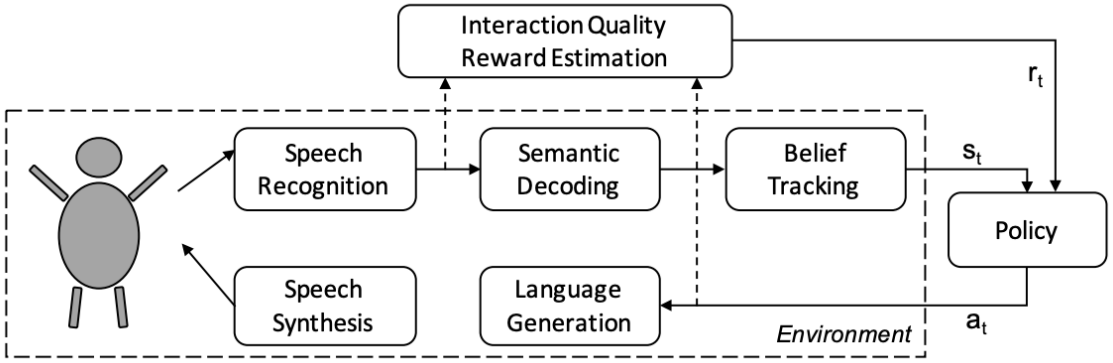
Learning with user satisfaction

Heinrich Heine
Universität Düsseldorf

- A data-driven method to evaluate the dialogue
- Identify problematic conversations
- Can we optimise the dialogue policy with user satisfaction reward estimation?
  - collecting end user ratings is not trivial
  - mapping questionnaire to a scalar reward value

Interaction quality reward estimation

- Interaction quality reward function can be used cross different corpus
  - independent of the user goal
  - independent of the domain information
- $R_{IQ} = T \cdot (-1) + (iq - 1) \cdot 5$
  - -1 :per turn penalty
  - iq: interaction quality (1-5)
  - T : max turn
- In comparison, $R_{TS} = T \cdot (-1) + \mathbf{1}_{TS} \cdot 20$
  - $\mathbf{1}_{TS}$: task success, 1 for success and 0 otherwise



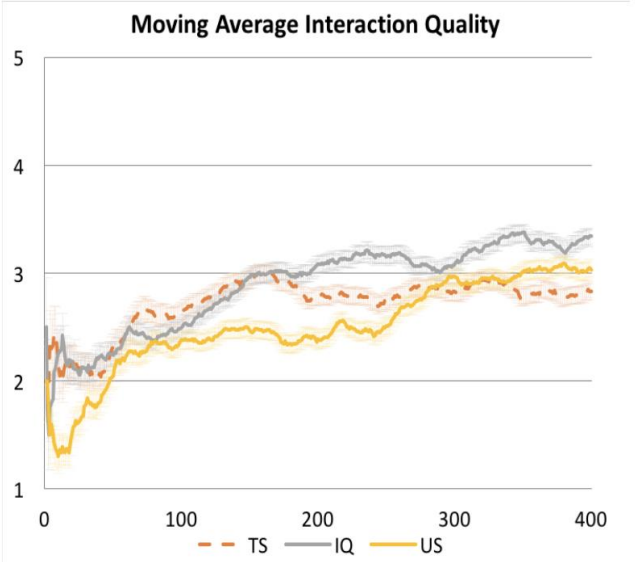
Simulation experiment

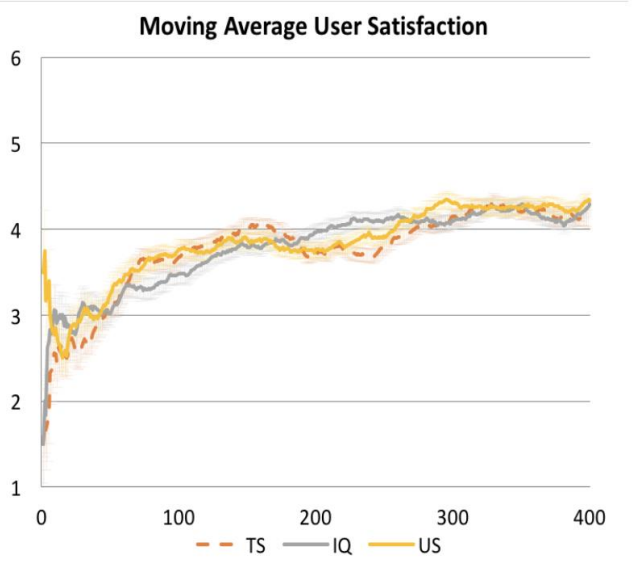
- Setup
  - IQ estimator is trained on LetsGo dataset
  - Train dialogue policy on five different corpus based on the GP-SARSA algorithm
- Task success rate (TSR)
  - The difference between source and target domain causes the different TSR
  - With the higher noise, the model should more focus on success
  - successful but noisy v.s not successful
- Average interaction quality (AIQ)
  - IQ-based model are better throughout the experiments

Domain	*SER*	*TSR*		*AIQ*		*ADL*	
		R_{TS}	R_{IQ}	R_{TS}	R_{IQ}	R_{TS}	R_{IQ}
CR	0%	0.98	0.98	3.88	3.96	4.37	4.34
	15%	0.86	0.85	3.51*	3.76*	5.21	4.93
	30%	0.84*	0.76*	3.34	3.46	5.73	5.54
CH	0%	0.97	0.96	3.02*	3.32*	5.74	5.79
	15%	0.79*	0.66*	2.69*	3.21*	7.27*	6.53*
	30%	0.62	0.55	2.13*	2.72*	8.81*	7.87*
SR	0%	0.93	0.93	2.88*	3.36*	6.31*	5.57*
	15%	0.58	0.65	2.5*	3.25*	8.03*	6.62*
	30%	0.46	0.41	2.17*	2.71*	9.13*	7.95*
SH	0%	0.94	0.93	3.1*	3.36*	5.66	5.92
	15%	0.71	0.67	2.61*	3.07*	7	6.73
	30%	0.51	0.5	2.29*	2.77*	8.94	8.64
L	0%	0.85	0.89	2.68*	3.11*	7.01*	6.15*
	15%	0.59	0.63	2.12*	2.97*	9.04*	6.72*
	30%	0.45	0.41	2.1*	2.52*	9.11*	8.09*
TV	0%	0.92*	0.86*	3.08*	3.42*	5.84	5.76
	15%	0.85*	0.78*	2.85*	3.44*	6.78*	5.88*
	30%	0.69	0.68	2.77*	3.06*	7.2	6.75

Learning from real humans

- Baseline
  - *subjective* task success: “Have you found all information you were looking for?” (1/0)
  - user satisfaction: “How satisfied are you with the interaction?” (1-6)
- Trained on CamRestaurant





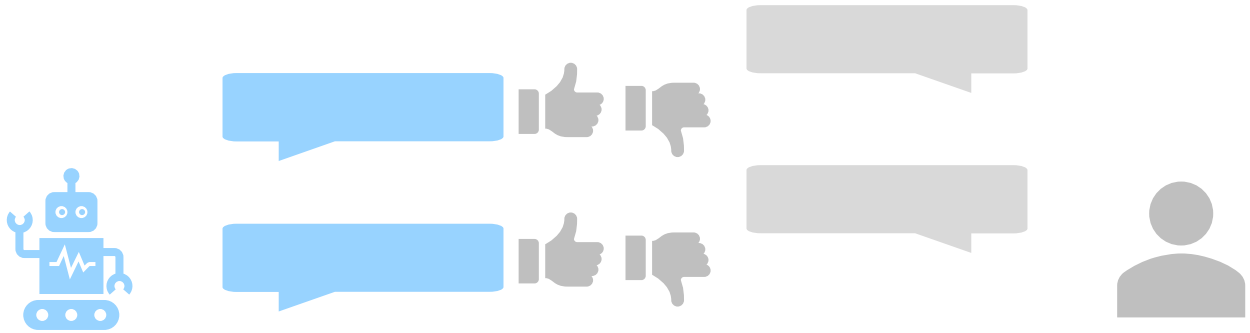
The challenge

Heinrich Heine
Universität Düsseldorf

- Labeling
- User satisfaction as the reward function

How to get labeled data

- Dialogue-level user satisfaction
  - The result from end users are noisy
  - It is hard to generalise because labeling by experts takes lots effort
- Turn-level interaction quality or response quality
  - Labeling by end users is interrupting and may casue dissatifaction
  - The annotation cost is higher than labeling the task success



The reward function is noisy

- Simulation training
  - The user simulator does not change its behaviour according to the satisfaction level
- Learning with real users
  - Pre-trained user satisfaction estimator as the reward function
    - Mismatch between the source domain and target domain
    - Influenced by the source system
  - Feedback from end users
    - Unreliable in usual
    - Uncertainty estimation, such as Gaussian process models (Su et all., 2016)
    - User persona learning

Reference

- Walker, M., Litman, D., Kamm, C. A., & Abella, A. (1997, July). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 271-280).
- Schmitt, A., & Ultes, S. (2015). Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, *74*, 12-36.
- Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019b. Multi-domain conversation quality evaluation via user satisfaction estimation. In 3rd Workshop on Conversation AI: Today's Practice and Tomorrow's Potential, 33rd Conference on Neural Information Processing Systems.
- Bodigutla, P. K., Tiwari, A., Matsoukas, S., Valls-Vargas, J., & Polymenakos, L. (2020, November). Joint Turn and Dialogue level User Satisfaction Estimation on Mulit-Domain Conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 3897-3909).

Reference

- Ultes, S., Budzianowski, P., Casanueva, I., Mrkšić, N., Rojas-Barahona, L., Su, P. H., ... & Young, S. (2017). Domain-Independent User Satisfaction Reward Estimation for Dialogue Policy Learning. *Proc. Interspeech 2017*, 1721-1725.
- Su, P. H., Gasic, M., Mrkšić, N., Barahona, L. M. R., Ultes, S., Vandyke, D., ... & Young, S. (2016, August). On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2431-2441).