User satisfaction in dialogue systems

Hsien-chin Lin
2021.05.14
Outline

- 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...

Key word extraction

Image classification

Dialogue system
User-centric dialogue system development

- 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?
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

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

Model overall rating on the dialogue level

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

User satisfaction model (Linear regression)
Problems for dialogue level user satisfaction

- **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

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

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

Interaction quality

Labeled by experts

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

![Diagram showing system interact with end users, log data, expert annotators, and user satisfaction model (SVM model).]
Interaction quality prediction

- **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

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

Problems for interaction quality

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

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
  - Cohesion between response and request
- ... Bodigutla et al., 2019b

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Turn Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>User: I would like to book a cab to work</td>
<td>5.0</td>
</tr>
<tr>
<td>System: Sure, at what time would you like to be picked up?</td>
<td></td>
</tr>
<tr>
<td>User: Eight a.m tomorrow</td>
<td>5.0</td>
</tr>
<tr>
<td>System: Could you confirm your cab booking to {office address} at 8 a.m tomorrow?</td>
<td></td>
</tr>
<tr>
<td>You will be charged $50 for this ride</td>
<td></td>
</tr>
<tr>
<td>User: No, cancel my ride</td>
<td></td>
</tr>
<tr>
<td>System: Ok, your booking is confirmed. Your card ending xxxx has been charged $50</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Joint turn and dialogue level user satisfaction

- A BiLSTM model optimises 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 utterance.
- The model puts more weight on the dissatisfactory turns according to the learnt attention weights.

Bodigutla et al., 2020
Learning with user satisfaction

- 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, \( R_{TS} = T \cdot (-1) + 1_{TS} \cdot 20 \)
- \(1_{TS}\): 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

<table>
<thead>
<tr>
<th>Domain</th>
<th>SER</th>
<th>( R_{TS} )</th>
<th>( R_{IQ} )</th>
<th>( R_{TS} )</th>
<th>( R_{IQ} )</th>
<th>( R_{TS} )</th>
<th>( R_{IQ} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR</td>
<td>0%</td>
<td>0.98</td>
<td>0.98</td>
<td>3.88</td>
<td>3.96</td>
<td>4.37</td>
<td>4.34</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.86</td>
<td>0.85</td>
<td>3.51*</td>
<td>3.76*</td>
<td>5.21</td>
<td>4.93</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.84*</td>
<td>0.76*</td>
<td>3.34</td>
<td>3.46</td>
<td>5.73</td>
<td>5.54</td>
</tr>
<tr>
<td>CH</td>
<td>0%</td>
<td>0.97</td>
<td>0.96</td>
<td>3.02*</td>
<td>3.32*</td>
<td>5.74</td>
<td>5.79</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.79*</td>
<td>0.66*</td>
<td>2.69*</td>
<td>3.21*</td>
<td>7.27*</td>
<td>6.53*</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.62</td>
<td>0.55</td>
<td>2.13*</td>
<td>2.72*</td>
<td>8.81*</td>
<td>7.87*</td>
</tr>
<tr>
<td>SR</td>
<td>0%</td>
<td>0.93</td>
<td>0.93</td>
<td>2.88*</td>
<td>3.36*</td>
<td>6.31*</td>
<td>5.57*</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.58</td>
<td>0.65</td>
<td>2.5*</td>
<td>3.25*</td>
<td>8.03*</td>
<td>6.62*</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.46</td>
<td>0.41</td>
<td>2.17*</td>
<td>2.71*</td>
<td>9.13*</td>
<td>7.95*</td>
</tr>
<tr>
<td>SH</td>
<td>0%</td>
<td>0.94</td>
<td>0.93</td>
<td>3.1*</td>
<td>3.36*</td>
<td>5.66</td>
<td>5.92</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.71</td>
<td>0.67</td>
<td>2.61*</td>
<td>3.07*</td>
<td>7.09*</td>
<td>6.73</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.51</td>
<td>0.5</td>
<td>2.29*</td>
<td>2.77*</td>
<td>8.94</td>
<td>8.64</td>
</tr>
<tr>
<td>L</td>
<td>0%</td>
<td>0.85</td>
<td>0.89</td>
<td>2.68*</td>
<td>3.11*</td>
<td>7.01*</td>
<td>6.15*</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.59</td>
<td>0.63</td>
<td>2.12*</td>
<td>2.97*</td>
<td>9.04*</td>
<td>6.72*</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.45</td>
<td>0.41</td>
<td>2.1*</td>
<td>2.52*</td>
<td>9.11*</td>
<td>8.09*</td>
</tr>
<tr>
<td>TV</td>
<td>0%</td>
<td>0.92*</td>
<td>0.86*</td>
<td>3.08*</td>
<td>3.42*</td>
<td>5.84</td>
<td>5.76</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.85*</td>
<td>0.78*</td>
<td>2.85*</td>
<td>3.44*</td>
<td>6.78*</td>
<td>5.88*</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.69</td>
<td>0.68</td>
<td>2.77*</td>
<td>3.06*</td>
<td>7.2</td>
<td>6.75</td>
</tr>
</tbody>
</table>
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

- 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 cause dissatisfaction
  - 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 al., 2016)
  - User persona learning
Reference


