EVALUATING CONVERSATIONAL RECOMMENDER SYSTEMS VIA USER SIMULATION

Shuo Zhang*
Bloomberg, London, UK
@imsure318

Krisztian Balog
University of Stavanger, Norway
@krisztianbalog

* Work done while at the University of Stavanger, Norway.
MOTIVATION

• Test-collection based ("offline") evaluation
  • Possible to create a reusable test collection for a specific subtask ☑
  • Limited to a single turn, does not measure overall user satisfaction ❌

• Human evaluation
  • Possible to annotate entire conversations ☑
  • Expensive, time-consuming, does not scale ❌

• Evaluation of conversational information access systems is an open challenge. We explore user simulation in this work.
OBJECTIVES

- Develop a user simulator that
  - produces responses that a real user would give in a certain dialog situation
  - enables automatic assessment of conversational agents
  - makes no assumptions about the inner workings of conversational agents
  - is data-driven (requires only a small annotated dialogue corpus)
PROBLEM STATEMENT

• For a given system $S$ and user population $U$, the goal of user simulation $U^*$ is to predict the performance of $S$ when used by $U$, denoted as $M(S,U)$

• For two systems $S_1$ and $S_2$, $U^*$ should be such that

  $\begin{align*}
  \text{if} \quad & M(S_1,U) < M(S_2,U) \\
  \text{then} \quad & M(S_1,U^*) < M(S_2,U^*)
  \end{align*}$
APPROACH
SIMULATION FRAMEWORK

- Natural language understanding (NLU)
- Natural language generation (NLG)
- Response generation
  - Preference model
  - Interaction model

Conversational agent

Simulated user
SIMULATION FRAMEWORK

Conversational agent

Natural language understanding (NLU)

Response generation
- Preference model
- Interaction model

Simulated user

Translating an agent utterance into a structured format
Determining the next user action based on the understanding of the agent’s utterance
SIMULATION FRAMEWORK

Conversational agent

Natural language understanding (NLU)

Natural language generation (NLG)

Response generation

- Preference model
- Interaction model

Turning a structured response representation into natural language
MODELING SIMULATED USERS

• Model dialogue as a Markov Decision Process
• Every MDP is formally described by a finite space $S$, a finite action set $A$, and transition probabilities $P$

  • Dialogue acts (or actions): task-specific intents that are being communicated in utterances
  • Dialogue state: the state of the dialogue manager is in. At each time step (dialogue turn) $t$, the dialogue manager is in a particular state $s_t$
  • Transition probabilities: the probability of transitioning from $s_t$ to $s_{t+1}$
AGENDA-BASED SIMULATION*

- The action agenda $A$ is a stack-like representation for user actions that is dynamically updated.
- The next user action is selected from the top of the agenda.
- Agenda updates are regarded as a sequence of pull or push operations.
  - Accomplished goal -> pull operation
  - Not accomplished -> push operation

## ACTION SPACE*

<table>
<thead>
<tr>
<th>Action</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disclose</td>
<td>I would to arrange a holiday in Italy</td>
</tr>
<tr>
<td>Reveal</td>
<td>Actually, we need to go on the 3rd of May in the evening</td>
</tr>
<tr>
<td>Inquire</td>
<td>What other regions in Europe are like that?</td>
</tr>
<tr>
<td>Navigate</td>
<td>Which one is the cheapest option?</td>
</tr>
<tr>
<td>Note</td>
<td>That hotel could be a possibility</td>
</tr>
<tr>
<td>Complete</td>
<td>Thanks for the help, bye</td>
</tr>
</tbody>
</table>

*Azzopardi et al. Conceptualizing Agent-human interactions during the conversational search process. CAIR 2018*
INTERACTION MODEL

• The interaction model defines how the agenda should be initialized ($A_0$) and updated ($A_t \Rightarrow A_{t+1}$)

• QRFA Model*
  • User: Query and Feedback
  • Agent: Request and Answer
  • QRFA are mapped to action space manually

• CIR6 Model

* Vakulenko et al. QRFA: A Data-Driven Model for Information Seeking Dialogues. ECIR 2019.
The preference model is meant to capture individual differences and personal tastes.

Preferences are represented as a set of attribute-value pairs.

- **Single Item Preference**
  - Check if \( i \in I_u \) an answer accordingly, and randomly decide preference
  - It offers limited consistency

- **Personal Knowledge Graph**
  - PKG has two types of nodes: items and attributes
  - Infers the rating for that attribute by considering the ratings of items that have that attribute

\[
 r_j = \frac{1}{|I_j|} \sum_{i \in I_j} r_i
\]

EXPERIMENTAL EVALUATION
EVALUATION ARCHITECTURE

Three existing conversational movie recommenders (A, B, C) are compared using both real (_human_user_) and simulated (_simulated_user_) users.

Real users: we invite crowdsourcing workers to interact with recommenders on Telegram, and use their dialogue records for initializing the simulated users.
EVALUATION ARCHITECTURE

• Simulated users
  • Preference model is initialized by sampling historical preferences of a real user from MovieLens data
  • Interaction model is trained based on behaviors of real human users
  • Both NLU and NLG use hand-crafted templates
CHARACTERISTICS OF CONVERSATIONS

• (RQ1) How well do our simulation techniques capture the characteristics of conversations?

<table>
<thead>
<tr>
<th>Method</th>
<th>AvgTurns</th>
<th>UserActRatio</th>
<th>DS-KL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>Real users</td>
<td>9.20</td>
<td>14.84</td>
<td>20.24</td>
</tr>
<tr>
<td>QRFA-Single</td>
<td>10.52</td>
<td>12.28</td>
<td>17.51</td>
</tr>
<tr>
<td>CIR6-Single</td>
<td>9.44</td>
<td>12.75</td>
<td>15.92</td>
</tr>
<tr>
<td>CIR6-PKG</td>
<td>6.16</td>
<td>9.87</td>
<td>10.56</td>
</tr>
</tbody>
</table>

• CIR6-PKG tends to have significantly shorter average conversation length, since it terminates the dialog as soon as the user finds a recommendation they like
PERFORMANCE PREDICTION

• (RQ2) How well do the relative ordering of systems according to some measure correlate when using real vs. simulated users?

<table>
<thead>
<tr>
<th>Method</th>
<th>Reward</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real users</td>
<td>A (8.88) &gt; B (7.56) &gt; C (6.04)</td>
<td>B (0.864) &gt; A (0.833) &gt; C (0.727)</td>
</tr>
<tr>
<td>QRFA-Single</td>
<td>A (8.04) &gt; B (7.41) &gt; C (6.30)</td>
<td>B (0.836) &gt; A (0.774) &gt; C (0.718)</td>
</tr>
<tr>
<td>CIR6-Single</td>
<td>A (8.64) &gt; B (8.28) &gt; C (6.01)</td>
<td>B (0.822) &gt; A (0.807) &gt; C (0.712)</td>
</tr>
<tr>
<td>CIR6-PKG</td>
<td>A (11.12) &gt; B (10.65) &gt; C (9.31)</td>
<td>A (0.870) &gt; B (0.847) &gt; C (0.784)</td>
</tr>
</tbody>
</table>

Performance of conversational agents using real vs. simulated users, in terms of Reward and Success Rate. We show the relative ordering of agents (A–C), with evaluation scores in parentheses.

• High correlation between automatic and human evaluations
REALISTICITY

- (RQ3) Do more sophisticated simulation approaches (i.e., more advanced interaction and preference modeling) lead to more realistic simulation?

<table>
<thead>
<tr>
<th>Method</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Win</td>
<td>Loss</td>
<td>Tie</td>
<td>Win</td>
</tr>
<tr>
<td>QRFA-Single</td>
<td>20</td>
<td>39</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>CIR6-Single</td>
<td>27</td>
<td>30</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>CIR6-PKG</td>
<td>22</td>
<td>39</td>
<td>14</td>
<td>27</td>
</tr>
</tbody>
</table>

- Our interaction model (CIR6) and personal knowledge graphs for preference modeling both bring improvements
FURTHER ANALYSIS

• We analyze the reasons when the crowd workers chose the real users, and classify them as follows

<table>
<thead>
<tr>
<th>Style</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realisticity</td>
<td>how realistic or human-sounding a dialog is</td>
</tr>
<tr>
<td>Engagement</td>
<td>involvement of the user in the conversation</td>
</tr>
<tr>
<td>Emotion</td>
<td>expressions of feelings or emotions</td>
</tr>
<tr>
<td>Content</td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td>user does not seem to understand the agent correctly</td>
</tr>
<tr>
<td>Grammar</td>
<td>language usage, including spelling and punctuation</td>
</tr>
<tr>
<td>Length</td>
<td>the length of reply</td>
</tr>
</tbody>
</table>
SUMMARY OF CONTRIBUTIONS

• A general framework for evaluating conversational recommender agents via simulation

• Interaction and preference models to better control the conversation flow and to ensure the consistency of responses given by the simulated user

• Experimental comparison of three conversational movie recommender agents, using both real and simulated users

• Analysis of comments collected from human evaluation, and identification of areas for future development
THANK YOU!

• Resources: https://github.com/iai-group/UserSimConvRec