Interactive Evaluation of Dialog Track at DSTC9

Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi

Introduction

- The goal of dialog systems research is to create dialog systems that can have a back-and-forth conversation with a real user and strive for user satisfaction
- Most research focuses on response generation on static datasets which neglects important properties of dialog
- We introduce the Interactive Evaluation of Dialog Track at DSTC9, which challenges dialog researchers to:
  - build knowledge-grounded response generation models
  - extend models to interactive settings with real users

33 submissions to sub-task 1 and 11 submissions to sub-task 2

DSTC9 Track Setup

Sub-Task 1: Static Evaluation
- Build knowledge-grounded response generation models
- Evaluate on the Topical-Chat corpus:
  - Automatic metrics: METEOR, BERTScore, USR
  - Human evaluation: turn-level evaluation on AMT throughout track

Sub-Task 2: Interactive Evaluation
- Extend response models to interactive settings
- Evaluate on DialPort, with real users:
  - Automatic metrics: FED
  - Human evaluation: on-going dialog-level evaluation on AMT
  - User feedback: user feedback provided by users on DialPort was periodically shared with the participants

Interactive Evaluation on DialPort

- Dialog systems were evaluated on the DialPort platform
- Users interact with a system and can provide feedback.
- Systems are connected via an API interface
- Real users are recruited via Facebook advertising.
- An advertising budget of $2500 USD resulted in 4561 dialogs (41,600 turns). There were 2960 dialogs with at least 4 turns (total of 38,488 turns). Cost of < $1USD/dialog.
- Real users provided feedback! 3829 feedbacks (9% of turns), with 544 corrections, 517 written feedback and 2776 likes/dislikes

Interactive evaluation is feasible
- Our track demonstrated the feasibility and effectiveness of interactive evaluation with real users
- Real users have good interactions and leave feedback without false incentives (e.g., payment)
- Cost of dialog (post-filtering) is less than $1USD

Interactive evaluation is better than static
- Differences between systems are more pronounced
  - Back-and-forth interaction is more challenging
  - Real users are testing the limits of systems
  - Conversation is truly open-domain – rather than being constrained to specific topics
- Real users interact because they want to
  - Length of conversation strongly correlates with human judgements of quality – users talk longer to better systems!
  - Users provide valuable feedback – additional signal for system developers

Better evaluation metrics are important
- Reference-free evaluation metrics (USR/FED) correlate better than METEOR/BERTScore
  - METEOR: 0.23 (static)
  - BERTScore: 0.22 (static)
  - USR: 0.35 (static)
  - FED: 0.49 (interactive)

DSTC9 Track Results

<table>
<thead>
<tr>
<th>Static Evaluation</th>
<th>System</th>
<th>Avg. Turns</th>
<th>FED</th>
<th>Human</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top systems close in performance</td>
<td>12.84</td>
<td>4.97</td>
<td>4.15</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Automatic metrics are poor predictors</td>
<td>13.87</td>
<td>4.79</td>
<td>4.14</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Large-scale pre-trained models outperform those trained on Topical-Chat</td>
<td>8.89</td>
<td>4.61</td>
<td>4.08</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Interactive Evaluation</td>
<td>9.36</td>
<td>4.68</td>
<td>4.05</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>More variance between systems</td>
<td>9.62</td>
<td>4.53</td>
<td>3.93</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Top systems use pipelines (generation, ranking, post-processing)</td>
<td><em>6</em></td>
<td>8.75</td>
<td>4.72</td>
<td>3.87</td>
<td>6</td>
</tr>
<tr>
<td>FED has system-level correlations with human rankings</td>
<td>8.51</td>
<td>4.41</td>
<td>3.85</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Average number of turns has system-level correlations</td>
<td>7.67</td>
<td>4.30</td>
<td>3.85</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>User feedback: user feedback provided by users on DialPort was periodically shared with the participants</td>
<td>6.53</td>
<td>4.64</td>
<td>3.83</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7.35</td>
<td>4.80</td>
<td>3.69</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5.80</td>
<td>3.69</td>
<td>3.60</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

Interactive evaluation is feasible
- Our track demonstrated the feasibility and effectiveness of interactive evaluation with real users
- Real users have good interactions and leave feedback without false incentives (e.g., payment)
- Cost of dialog (post-filtering) is less than $1USD

Interactive evaluation is better than static
- Differences between systems are more pronounced
  - Back-and-forth interaction is more challenging
  - Real users are testing the limits of systems
  - Conversation is truly open-domain – rather than being constrained to specific topics
- Real users interact because they want to
  - Length of conversation strongly correlates with human judgements of quality – users talk longer to better systems!
  - Users provide valuable feedback – additional signal for system developers

Better evaluation metrics are important
- Reference-free evaluation metrics (USR/FED) correlate better than METEOR/BERTScore
  - METEOR: 0.23 (static)
  - BERTScore: 0.22 (static)
  - USR: 0.35 (static)
  - FED: 0.49 (interactive)