Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task

Sijie Tao and Tetsuya Sakai
Waseda University
Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions
Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions
History of the task

• NTCIR-14, Jun 2019, Short Text Conversation Task (STC-3) [Zeng+19]
  • DCH-1 Dataset used as training and test sets
  • 3,700 + 390 for Chinese, 1,672 + 390 for English

• NTCIR-15, Dec 2020, Dialogue Evaluation Task (DialEval-1) [Zeng+20]
  • DCH-1 used as training and development sets, new test set built
  • 3,700 + 390 + 300 for Chinese, 2,251 + 390 + 300 for English

• NTCIR-16, Jun 2022, Dialogue Evaluation Task (DialEval-2)
  • DCH-2 [Zeng+21] as training and development sets, new test set built
  • 4,090 + 300 + 65 for both Chinese and English
Task Definition

• DialEval-2 hosts two subtasks:
  • Dialogue Quality (**DQ**)  
  • Nugget Detection (**ND**)  

• **DQ**: Given a customer-helpdesk dialogue, return an estimated distribution of dialogue quality ratings for the entire dialogue  

• **ND**: Given a customer-helpdesk dialogue, return an estimated distribution of labels over nugget types for each turn
An Example of a customer-helpdesk dialogue [Zeng+20]

C: The Smartisan App Store of my mobile phone has been disabled for nearly half a month and the system couldn't be updated. The network was normal. Please give me an explanation.
2016-5-22 13:45

H: To ensure information security, we updated the system security encryption algorithm. Please visit the website, and download and install "System Update Service" to update your system. For detailed operations, please visit the link.
2016-5-22 13:56

C: It worked properly. Thank you!
2016-5-22 23:40

H: You are welcome.
2016-5-22 23:50
Dialogue Quality Subtask (DQ)

• Given a customer-helpdesk dialogue, return an estimated distribution of dialogue quality ratings for the entire dialogue.

• Three types of dialogue quality ratings (Likert scale -2 to 2):
  • A-score: Task Accomplishment
  • S-score: Customer Satisfaction (about the dialogue itself, not about the product/service)
  • E-score: Dialogue Effectiveness
Dialogue Quality Subtask (DQ)

Annotators

Customer’s turn
Helpdesk’s turn
Customer’s turn
Helpdesk’s turn

Gold distribution

Distribution estimated by System X

Distribution estimated by System Y

X is better than Y
Dialogue Quality Subtask (DQ)

- Evaluation metrics
  - NMD (Normalised Match Distance)
  - RSNOD (Root Symmetric Normalised Order-aware Divergence) [Sakai18]

- Both measures take into account the distance between two bins, to make sure X is rated higher than Y in the previous slide.
Nugget Detection Subtask (ND)

• What is a nugget?
An Example of a customer-helpdesk dialogue [Zeng+20]

C: The Smartisan App Store of my mobile phone has been disabled for nearly half a month and the system couldn't be updated. The network was normal. Please give me an explanation.
2016-5-22 13:45

H: To ensure information security, we updated the system security encryption algorithm. Please visit the website, and download and install "System Update Service" to update your system. For detailed operations, please visit the link.
2016-5-22 13:56

C: It worked properly. Thank you!
2016-5-22 23:40

H: You are welcome
2016-5-22 23:50
Nugget Detection Subtask (ND)

• What is a nugget?

Initial state (Customer facing a problem)

Intermediate state

Goal state (problem Solved)

Turn that does not serve as a nugget

Nugget = a turn by Customer or Helpdesk that helps Customer transition towards the goal ("problem solved") state
Nugget Detection Subtask (ND)

- Given a customer-helpdesk dialogue, return an estimated distribution of labels over nugget types for each turn

<table>
<thead>
<tr>
<th>Nugget type</th>
<th>Customer</th>
<th>Helpdesk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigger</td>
<td>CNUG0: tell the problem to Helpdesk</td>
<td></td>
</tr>
<tr>
<td>Regular</td>
<td>CNUG</td>
<td>HNUG</td>
</tr>
<tr>
<td>Goal</td>
<td>CNUG*: tell Helpdesk that the problem has been solved</td>
<td>HNUG*: tell Customer the solution to the problem</td>
</tr>
<tr>
<td>Not-a-nugget</td>
<td>CNaN</td>
<td>HNaN</td>
</tr>
</tbody>
</table>
Nugget Detection Subtask (ND)

Gold distributions

Customer’s turn
Helpdesk’s turn
Customer’s turn
Helpdesk’s turn

Distributions estimated by System X
Nugget Detection Subtask (ND)

• Evaluation metrics
  • RNSS (Root Normalised Sum of Squares)
  • JSD (Jensen-Shannon Divergence) [Sakai18]

• No need to use NMD or RSNOD, as the bins in the ND subtask are nominal (e.g. HNUG, HNUG*, HNaN), not ordinal
Motivation of the task

• Evaluate customer-helpdesk dialogues automatically

• DQ: An effective DQ system is useful for building helpdesk systems that can generate effective utterances for diverse users.

• ND: An effective ND system is useful for building effective helpdesk systems that can self diagnose at the dialogue turn level to improve themselves.
Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions
The new data collection for DialEval-2

• For DialEval-2, we use DCH-2 dataset [Zeng+21] as training and development sets
• A new test set which contains 65 dialogues is additionally built

<table>
<thead>
<tr>
<th>Source</th>
<th>Training</th>
<th>Chinese</th>
<th>Test</th>
<th>Training</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DCH-2</td>
<td>DCH-2</td>
<td>Weibo</td>
<td>Translation</td>
<td></td>
</tr>
<tr>
<td>#dialogues</td>
<td>4,090</td>
<td>300</td>
<td>65</td>
<td>4,090</td>
<td>300</td>
</tr>
<tr>
<td>#annotators/dialogue</td>
<td>19</td>
<td>20</td>
<td>20</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Quality annotation criteria</td>
<td>A-score, E-score, S-score (See Section 2.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget types</td>
<td>CNUG0, CNUG, HNUG, CNUG*, HNUG* (See Section 2.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions
Participant teams (only four, last time we had seven)

- IMNTPU (National Taipei University) [Hsiao+22]
- NKUST (National Kaohsiung University of Science and Technology) [Chang+22]
- RSLDE (Waseda University) [Li+22]
- TUA1 (Tokushima University) [Ding+22]

<table>
<thead>
<tr>
<th>Teams</th>
<th>Runs</th>
<th>Chinese</th>
<th></th>
<th>English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DQ</td>
<td>ND</td>
<td>DQ</td>
<td>ND</td>
</tr>
<tr>
<td>IMNTPU</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NKUST</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>RSLDE</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>TUA1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>9</strong></td>
<td><strong>7</strong></td>
<td><strong>7</strong></td>
<td><strong>4</strong></td>
<td><strong>6</strong></td>
</tr>
</tbody>
</table>
Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions
Results

• Baselines (exactly the same as the baselines in DialEval-1) [Zeng+20]
  • BL-lstm (Baseline-run0): A baseline model which leverages Bidirectional Long Short-term Memory;
  • BL-uniform (Baseline-run1): A baseline model which always predict the uniform distribution;
  • BL-popularity (Baseline-run2): A baseline model which predicts the probability of the most popular label as one, and predicts other labels as 0.
• **TUA1-run1, 2** are the top runs in terms of RSNOD and NMD for A and S-score

• Only **TUA-run0** outperforms Baseline-run0 statistically significantly in terms of NMD for E-score
• **TUA1-run0** is the top run and the only run that outperforms the baseline systems
• But the differences between **TUA1-run0** and the top baselines are not statistically significant
Results (ND, Chinese)

- **RSLDE-run0** is the top run and the only run that can outperform Baseline-run0 in terms of both JSD and RNSS.
- But the difference between them is not statistically significant.

### Table 7: Chinese Nugget Detection Results

<table>
<thead>
<tr>
<th>Run</th>
<th>Mean JSD</th>
<th>Run</th>
<th>Mean RNSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSLDE-run0</td>
<td>0.0560</td>
<td>RSLDE-run0</td>
<td>0.1604</td>
</tr>
<tr>
<td>Baseline-run0</td>
<td>0.0585</td>
<td>Baseline-run0</td>
<td>0.1651</td>
</tr>
<tr>
<td>RSLDE-run2</td>
<td>0.0607</td>
<td>RSLDE-run1</td>
<td>0.1712</td>
</tr>
<tr>
<td>RSLDE-run1</td>
<td>0.0634</td>
<td>RSLDE-run2</td>
<td>0.1720</td>
</tr>
<tr>
<td>NKUST-run0</td>
<td>0.0670</td>
<td>NKUST-run0</td>
<td>0.1761</td>
</tr>
<tr>
<td>TUA1-run0</td>
<td>0.0700</td>
<td>TUA1-run0</td>
<td>0.1780</td>
</tr>
<tr>
<td>Baseline-run2</td>
<td>0.1864</td>
<td>Baseline-run2</td>
<td>0.2901</td>
</tr>
<tr>
<td>Baseline-run1</td>
<td>0.2042</td>
<td>Baseline-run1</td>
<td>0.3371</td>
</tr>
<tr>
<td>NKUST-run1</td>
<td>0.2432</td>
<td>NKUST-run1</td>
<td>0.3774</td>
</tr>
<tr>
<td>TUA1-run1</td>
<td>0.2909</td>
<td>TUA1-run1</td>
<td>0.3939</td>
</tr>
</tbody>
</table>
Results (ND, English)

- RSLDE-run0 and IMNTPU-run0 are the runs that can outperform Baseline-run0.
- But their differences between the baseline are not statistically significant.

<table>
<thead>
<tr>
<th>Run</th>
<th>Mean JSD</th>
<th>Run</th>
<th>Mean RNSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSLDE-run0</td>
<td>0.0557</td>
<td>IMNTPU-run0</td>
<td>0.1574</td>
</tr>
<tr>
<td>IMNTPU-run0</td>
<td>0.0601</td>
<td>RSLDE-run0</td>
<td>0.1615</td>
</tr>
<tr>
<td>Baseline-run0</td>
<td>0.0625</td>
<td>Baseline-run0</td>
<td>0.1722</td>
</tr>
<tr>
<td>NKUST-run0</td>
<td>0.0641</td>
<td>NKUST-run0</td>
<td>0.1744</td>
</tr>
<tr>
<td>RSLDE-run2</td>
<td>0.0676</td>
<td>RSLDE-run2</td>
<td>0.1778</td>
</tr>
<tr>
<td>RSLDE-run1</td>
<td>0.0691</td>
<td>TUA1-run0</td>
<td>0.1830</td>
</tr>
<tr>
<td>TUA1-run0</td>
<td>0.0728</td>
<td>RSLDE-run1</td>
<td>0.1853</td>
</tr>
<tr>
<td>Baseline-run2</td>
<td>0.1864</td>
<td>Baseline-run2</td>
<td>0.2901</td>
</tr>
<tr>
<td>Baseline-run1</td>
<td>0.2042</td>
<td>Baseline-run1</td>
<td>0.3371</td>
</tr>
</tbody>
</table>
Results
(Differences between metrics)

- The difference between different metrics are not statistically significant for both ND and DQ subtasks
- Consistent with what we observed at DialEval-1 and STC-3. [Zeng+19][Zeng+20]

<table>
<thead>
<tr>
<th>Table 8: Ranking Correlation between of Chinese runs ranked by two different metrics (Kendall’s τ with 95% CIs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue Quality (A-score)</td>
</tr>
<tr>
<td>NMD vs RSNOD</td>
</tr>
<tr>
<td>Dialogue Quality (S-score)</td>
</tr>
<tr>
<td>NMD vs RSNOD</td>
</tr>
<tr>
<td>Dialogue Quality (E-score)</td>
</tr>
<tr>
<td>NMD vs RSNOD</td>
</tr>
<tr>
<td>Nugget Detection</td>
</tr>
<tr>
<td>JSD vs RNSS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 13: Ranking Correlation between of English runs ranked by two different metrics (Kendall’s τ with 95% CIs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue Quality (A-score)</td>
</tr>
<tr>
<td>NMD vs RSNOD</td>
</tr>
<tr>
<td>Dialogue Quality (S-score)</td>
</tr>
<tr>
<td>NMD vs RSNOD</td>
</tr>
<tr>
<td>Dialogue Quality (E-score)</td>
</tr>
<tr>
<td>NMD vs RSNOD</td>
</tr>
<tr>
<td>Nugget Detection</td>
</tr>
<tr>
<td>JSD vs RNSS</td>
</tr>
</tbody>
</table>
Outline

1. History, definition, and motivation of DialEval
2. The new data collection for DialEval-2
3. Participants
4. Results
5. Conclusions
Conclusions

• Overview of DialEval-2:
  • Task definition
  • Data collection
  • Evaluation results

• From the evaluation results, we observe that
  • Only one run from TUA1 outperform the LSTM baseline significantly in Chinese DQ task in terms of NMD for E-score.
  • In other subtasks, none of the runs can outperform the LSTM baseline significantly.
  • No substantial difference is observed between the evaluation metrics for each subtasks.
References


References


