Improving the Performance of Dialogue Systems via Dialogue Breakdown Detection

Michimasa Inaba and Kenichi Takahashi
Hiroshima City University
Background

• Dialogue breakdown detection
  ■ A technique for identifying inappropriate utterances in dialogue systems
  ■ Particularly in chat-oriented systems

• Dialogue Breakdown Detection Challenge (DBDC)
  ■ International competitions focused on the performance of this technique
  ■ The third competition was held as one of the tasks of DSTC 6 in December 2017
Significance of Dialogue Breakdown Detection

The technique enables systems to avoid the creation of inappropriate utterances and also identify dialogue breakdowns when they occur [Higashinaka+ 16]

Does that mean that system performance will be improved by using it?

Specifically, how do I apply it to the system?

Can it truly avoid the dialogue breakdown?
Purpose

• Apply the dialogue breakdown detection technique to response selections in chat-oriented dialogue systems
  
  - Propose three application methods for using the technique
  - Experimentally verify whether the response performance is improved

![Diagram]

Chat-oriented dialogue system + Dialogue Breakdown Detection = Improved dialogue system?
Approach

• **Re-rank response candidates** by applying dialogue breakdown detection
  - Systems that can generate multiple responses often score their candidates
  - Alter the scores using the results of the breakdown detector

• **Analyze changes in performance**
Dialogue Systems in this Study

- **Example-based system (IRS)**
  - Based on IR-STATUS [Ritter+ 11]
  - Used in DBDC for constructing dataset

- **Neural Conversational Model (NCM)**
  - Dialogue system with an encoder-decoder neural network for response generation

- **Neural Utterance Ranking Model (NUR)**
  - Based on our neural network-based dialogue model [Inaba+ SIGDIAL2016]
Annotated Dialogue Data

- Dialogue rules and data format are fully compliant with DBDC
  - Dialogue data comprises Japanese text chats between a system and a user
  - Three types of breakdown labels are annotated by multiple annotators

**NB (Not a Breakdown):**
- It is easy to continue the conversation

**PB (Possible Breakdown):**
- It is difficult to smoothly continue the conversation

**B (Breakdown):**
- It is difficult to continue the conversation
S: Hello. How have you been recently?
U: My sinuses are really bothering me.
S: Sinus victims feel terrible.
U: That’s right. I can’t stop sneezing and have a runny nose.
S: A sneezing face looks so terrible.
U: So a mask is an indispensable item.
S: The loud sound of my sneezes is more worrying.
U: I agree that the sound can’t be stopped by a mask.
S: Sneezing is bad for my health
## Data Statistics

<table>
<thead>
<tr>
<th></th>
<th>IRS</th>
<th>NCM</th>
<th>NUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Dialogues</td>
<td>100</td>
<td>100</td>
<td>120</td>
</tr>
<tr>
<td>Number of Annotators</td>
<td>30</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>Fleiss’ κ</td>
<td>0.29</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>NB (Not a Breakdown)</td>
<td>31.1%</td>
<td>47.4%</td>
<td>57.7%</td>
</tr>
<tr>
<td>PB (Possible Breakdown)</td>
<td>26.7%</td>
<td>32.7%</td>
<td>27.0%</td>
</tr>
<tr>
<td>B (Breakdown)</td>
<td>42.1%</td>
<td>17.2%</td>
<td>15.2%</td>
</tr>
</tbody>
</table>

- System performance (based on NB labels)
  - NUR > NCM > IRS
- Data is used for training a dialogue breakdown detector
Dialogue Breakdown Detector

• **Best performance detector** from DBDC2 [Sugiyama 16]
  - Using typical error patterns of system responses as features
    - Unnatural connections of dialogue acts
    - Abrupt changes in the discussed topic etc.
  - Outputs probability distribution of labels $(P(NB), P(PB), P(B))$ for an input system’s response
Re-Ranking Methods
Re-Ranking Method (1/3)

- **Classification-based method**
  - Focus on the classification results
  - Lower the ranking of response candidates classified as causing a dialogue breakdown
  - Labels with the maximum probability are considered as classification results

<table>
<thead>
<tr>
<th>Before</th>
<th>Detection Result</th>
<th>Re-Ranked</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Candidate A</td>
<td>1st</td>
</tr>
<tr>
<td>2nd</td>
<td>Candidate B</td>
<td>2nd</td>
</tr>
<tr>
<td>3rd</td>
<td>Candidate C</td>
<td>3rd</td>
</tr>
</tbody>
</table>

Before Detection Result

Breakdown

NB

Candidate B

Candidate C

Candidate A
Re-Ranking Method (2/3)

• Probability-based method

- Calculate new score $s_{new}$ for re-ranking

$$s_{new} = sP(\text{non-breakdown})$$

$s$ is the response score

$$P(\text{non-breakdown}) = P(\text{NB}) \text{ or } P(\text{NB}) + P(\text{PB})$$

<table>
<thead>
<tr>
<th>Before</th>
<th>$s$</th>
<th>$P(\text{non-breakdown})$</th>
<th>Re-Ranked</th>
<th>$s_{new}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>A</td>
<td>2.0</td>
<td>1st</td>
<td>B</td>
</tr>
<tr>
<td>2nd</td>
<td>B</td>
<td>1.0</td>
<td>2nd</td>
<td>A</td>
</tr>
<tr>
<td>3rd</td>
<td>C</td>
<td>0.4</td>
<td>3rd</td>
<td>C</td>
</tr>
</tbody>
</table>

Candidate A  | 2.0 | 0.3 | 1st | Candidate B | 0.8 |
Candidate B  | 1.0 | 0.8 | 2nd | Candidate A | 0.6 |
Candidate C  | 0.4 | 0.5 | 3rd | Candidate C | 0.2 |
Re-Ranking Method (3/3)

- **Regression-based method**
  - Calculate the new score using the linear regression model
  - Use estimated probabilities $P(NB)$, $P(PB)$, $P(B)$ and the response score $s$ as input features
  - Evaluate using 10-fold cross-validation
    - This method includes the optimization of parameters
Experiment
Re-ranking Experiment

• **Experiment objectives**
  1. To verify the performance improvement by re-ranking given candidate responses
  2. To identify the application method that is effective

• **Data**
  - Train data for the dialogue breakdown detector (previously explained)
  - Evaluation data for re-ranking
Evaluation Data for Re-ranking

- Including context/responses pairs
  - Context
    - Chat log between two speakers
  - Responses
    - 10 or more response candidates with scores generated by the system
- Response candidates are annotated using breakdown labels (NB, PB, and B)
- Constructed 300 pairs for each system
A: My motivation to attend college is superficial.
B: I see. It’s dawn, and although I’m downhearted now, my motivation for working is not.
A: Sure.
B: Indeed! I am highly motivated on an average day, but once I get depressed, things become increasingly worse.
A: Oh, do not be so depressed and cheer up!

Annotations

NB 0.8 It requires a considerable degree of motivation.
PB 0.6 Self-education makes it difficult for me.
B 0.5 What is my motivation to ride a bicycle?
PB 0.2 It has to be a motivation.
...
Evaluation

• Re-ranking results are evaluated using Mean Average Precision (MAP)
  ■ Indicates the number of NB responses that are ranked high
  ■ Gives values between 0 and 1, with higher values indicating more accurate ranking predictions
Experimental Result
Result: IRS (Example-Based)

- Large improvements were observed compared to results without re-ranking
- Probability-based (NB) method was the most effective
Result: NCM (Neural-Based)

- Regression-based method was the most effective
- Performance improvement range is smaller than IRS

Classification-based methods showed deterioration compared to results without re-ranking
Result: NUR (Our System)

- Regression-based method was the most effective
- Performance improvement range was the smallest
Experiment: Summary

• Probability-based and regression-based methods are effective

• Classification-based method is unstable
  ⇒ Probability distribution is more important than classification result

• Effectiveness differed from system to system
  ⇒ A system with low performance has room for improvement
Discussion

• Analyzing changes in responses
  ▪ Comparing top-ranked responses before and after re-ranking

Number of tokens and types in a response

<table>
<thead>
<tr>
<th></th>
<th>Tokens (before)</th>
<th>Tokens (re-ranked)</th>
<th>Types (before)</th>
<th>Types (re-ranked)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRS</td>
<td>17.24</td>
<td>11.16</td>
<td>1219</td>
<td>761</td>
</tr>
<tr>
<td>NCM</td>
<td>10.13</td>
<td>8.47</td>
<td>363</td>
<td>332</td>
</tr>
<tr>
<td>NUR</td>
<td>9.69</td>
<td>9.54</td>
<td>885</td>
<td>881</td>
</tr>
</tbody>
</table>

Decreased both the number of tokens and types
Changes in Response (IRS)

Context:
Russian coke tastes good!

Response before re-ranking:
Violin! It is a waste that you practiced for eight years, but you cannot play it now. Why don't you start learning it again?

Re-ranked:
That's delicious.

Many IRS responses are changed to short and simple responses
Changes in Response (NCM)

Context :  
  I did not do my homework, but I came to class.

Response before re-ranking :  
  That's right, you don’t have any money.

Re-ranked :  
  Really?
Changes in Response (NUR)

Context:
A: Recent Lexus’s TV commercial is funny!
B: I don't like Lexus’s face.
A: How about Mazda’s?

Response before re-ranking:
I cannot say that Lexus is a fun car.

Re-ranked:
I am really looking forward to Mazda’s sports concept car.
Conclusion

• Applied a dialogue breakdown detection technique to 3 different types of systems
• Proposed three types of response re-ranking methods for applying the dialogue breakdown detector
• Experimental results indicate that breakdown probabilities are effective for performance improvement
• The disadvantage is that simple and dull responses increased