RSLDE at the NTCIR-16 Dialogue Evaluation
Fan Li, Tetsuya Sakai
Waseda University

Abstract
- RSLDE at DialEval-2
- English and Chinese Dialogue Quality (DQ) subtask
- English and Chinese Nugget Detection (ND) subtask
- Key Challenges
- DQ task: Representation of the structure of dialogue
- ND task: Representation of Dialogue Structure
- Does Dialogue Context Matter?

Introduction

The RSLDE team participated in the English and Chinese dialogue quality (DQ) and nugget detection (ND) subtasks of DialEval-2 [1]. Our proposed model:

- For ND Task:
  - Sentence Level Model based on BERT [2] and XLNet [3]
  - Dialogue Level Model: Dialogue Embedding + Transformer Encoder + Feed-Forward
- For DQ Task:
  - Dialogue Level Model: Dialogue Embedding + Transformer Encoder + Feed-Forward
  - Dialogue Level Baseline: Dialogue Embedding + Feed-Forward

Idea: Nugget has pattern of words

- CNUQ:
  - "Of Ctrip Customer Service Please read the details in the picture."
  - "Of Smartisan Technology Customer Service ... the wireless ring of mobile phone?"
- CNUQ+:
  - "Thank you."
  - "A customer service staff called to explain the problem this morning. I'm satisfied with this reply. The staff's attitude was sincere. I think Unicom is quite good."
- LNUQ:
  - "We will rework our efforts to do better service."
  - "You're welcome. That's our job."

Sentence Level Model

Language Model
BERT and XLNet (Permutation AR)

\[ \text{AR:} \quad \text{BERT :} \quad \text{Interval Segment Embedding} \]

\[ \text{PermutationAR} \quad \text{BERT} \quad \text{Interval Segment Embedding} \]

Figure 1: Comparison between BERT and XLNet

Approach

- Split the dialogues by turn, and each turn contains one or more utterances from either a customer or a helpdesk.
- Rebuild two datasets that contain only utterances of the customer and helpdesk respectively.
- Feed these 2 datasets into sentence-level model.

Dialogue Model

- Insert [CLS] and [SEP] tokens at the beginning and end of each utterance.
- Modify BERT’s interval segment embeddings.
- \( E_x \) represents customer’s utterance and \( E_y \) represents helpdesk’s utterance.
- Thus, the model knows where this utterance comes from.

Dialogue Model Structure

Experiments

Loss Function (for all tasks): \( \text{loss} = \sum_{i=1}^{m} \log \hat{y}_i \)

Metrics:
- Dialogue Quality: Normalised Match Distance
- Nugget Detection: Jensen-Shannon Divergence

Submitted Runs:

<table>
<thead>
<tr>
<th>Task</th>
<th>Language</th>
<th>Run</th>
<th>Model</th>
<th>Batch size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ</td>
<td>English</td>
<td>0</td>
<td>Dialogue-BERT</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>1</td>
<td>Dialogue-BERT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>3</td>
<td>Dialogue-BERT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>5</td>
<td>Dialogue-BERT</td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td>English</td>
<td>7</td>
<td>Dialogue-BERT</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>9</td>
<td>Dialogue-BERT</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: English A-score

Table 1: English B-score

Table 1: English C-score

Table 1: English D-score

Table 1: English E-score

Table 2: Chinese A-score

Table 2: Chinese B-score

Table 2: Chinese C-score

Table 2: Chinese D-score

Table 2: Chinese E-score

Conclusion

- The XLNet model has an outstanding language understanding capability for customer-helpdesk dialogues.
- Considering the structure and context information of a dialogue is important for the dialogue nugget detection.

References