Evaluating Dialogue Generation Systems via Response Selection

Shiki Sato¹  Reina Akama¹,²  Hiroki Ouchi²,¹  Jun Suzuki¹,²  Kentaro Inui¹,²

¹Tohoku University  ²RIKEN

{shiki.sato,reina.a,jun.Suzuki,inui}@ecei.tohoku.ac.jp
hiroki.ouchi@riken.jp

Our test set available at https://github.com/cl-tohoku/eval-via-selection
Overview

Motivation
• With a high correlation with humans
• At low cost

Approach
Response Selection (RS) with well-chosen false candidates

Context Candidates
How is he? A. He is fine.
B. She is fine
C. Is he fine? similar to GOLD but unacceptable

Contributions
1. Development of a RS test set with well-chosen candidates
2. Our comparison method correlates with human judgements
Necessity of Automatic Evaluation

DGSs can be compared by human evaluation

<table>
<thead>
<tr>
<th>Context</th>
<th>DGSs</th>
<th>Generated Responses</th>
<th>evaluate responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>How is he?</td>
<td>A</td>
<td>He is fine.</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>She is fine.</td>
<td>✗</td>
</tr>
</tbody>
</table>

Cons

Takes high cost

Cannot evaluate a lot of DGSs
Necessity of Automatic Evaluation

Automatic evaluation drives research as preliminary evaluation

Develop DGSs → Automatic Evaluation → Human Evaluation → Deploy
Comparing DGSs via RS

Method
We focus on comparing DGSs by RS accuracy

<table>
<thead>
<tr>
<th>Context</th>
<th>Candidates</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>How is he?</td>
<td>A. He is fine</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>B. She is fine</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>C. Is he fine?</td>
<td>20</td>
</tr>
</tbody>
</table>

Pros
RS can remedy one-to-many problem
Comparing DGSs via RS

Method
We focus on comparing DGSs by RS accuracy

<table>
<thead>
<tr>
<th>Context</th>
<th>Candidates</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>How is he?</td>
<td>A. He is fine</td>
<td>40</td>
</tr>
<tr>
<td>DGS A</td>
<td>B. She is fine</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>C. Is he fine?</td>
<td>20</td>
</tr>
</tbody>
</table>

select candidates with the lowest perplexity

Pros
RS can remedy one-to-many problem
Comparing DGSs via RS

Method
We focus on comparing DGSs by RS accuracy

Pros
RS can remedy one-to-many problem
Problems on RS false candidates

Undesirable false candidates are sampled by random-sampling

<table>
<thead>
<tr>
<th>Context</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>How is he?</td>
<td>A. He is fine.</td>
</tr>
<tr>
<td>randomly sampled</td>
<td>B. That is a car</td>
</tr>
<tr>
<td></td>
<td>C. I don’t know</td>
</tr>
</tbody>
</table>
Problems on RS false candidates

Undesirable false candidates are sampled by *random-sampling*

<table>
<thead>
<tr>
<th>Context</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>How is he?</td>
<td>① He is fine.</td>
</tr>
<tr>
<td>randomly sampled</td>
<td>B. That is a car</td>
</tr>
<tr>
<td></td>
<td>C. I don’t know</td>
</tr>
</tbody>
</table>

(i) Containing unrelated words
Problems on RS false candidates

Undesirable false candidates are sampled by random-sampling

Context

How is he? randomly sampled

Candidates

A. He is fine.
B. That is a car
C. I don’t know

(i) Containing unrelated words
(ii) Acceptable response
Contributions

1. Development of a RS test set with **well-chosen** false candidates

   “well-chosen”?
   - Similar to the GOLD
   - But unacceptable

2. Our comparison method correlates with human judgements
Test Set Construction Method

Collect false candidates in two steps:

1. Retrieve utterances similar to GOLD
   - Context: How is he?
   - GOLD: He is fine
   - Query: He is fine
   - Collect hardly distinguishable candidates
   - Human evaluation

2. Filter out acceptable utterances
   - Retrieved Utterances:
     - She is fine
     - Is he fine?
     - I don’t know
   - Retain only unacceptable candidates
Example of Our RS Test Set

Context
A: Excuse me. Could you please take a picture of us with this camera?
B: Sure. Which button do I press to shoot?
A: This one.

Chosen Candidates
1. Do I have to focus it?
2. But I do have ninja focus.
3. Do not lose your focus!
4. Could he not focus on that?

Containing “focus” related to “camera”

1,019 questions available at:
https://github.com/cl-tohoku/eval-via-selection
Experiments

Can our method compare DGSs like humans?

Experimental Procedure

1. Train 10 DGSs

2. Rank 10 DGSs

3. Compute rank correlation

DGSs

GRU [Cho+‘14], LSTM [Hochreiter+’97], Transformer [Vaswani+’17]
## Results: Correlation between DGS Rankings

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Spearman</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>-0.36</td>
<td>0.30</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.085</td>
<td>0.82</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.073</td>
<td>0.84</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>Our Method</td>
<td>0.48</td>
<td>0.19</td>
</tr>
</tbody>
</table>

*(word overwrap-based metrics)*
Results: Correlation between DGS Rankings

Correlation with ranking of DGSs by humans

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Spearman</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>-0.36</td>
<td>0.30</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.085</td>
<td>0.82</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.073</td>
<td>0.84</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>Our Method</td>
<td>0.48</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Correlates with humans more strongly than word overwrap-based metrics.
Conclusion

Motivation
• With high correlation with humans
• At low cost

Approach
Response Selection with well-chosen false candidates

Results
Our comparison method correlates with human judgements

Thank You for Listening!

shiki.sato@ecei.tohoku.ac.jp

Our test set available at https://github.com/cl-tohoku/eval-via-selection