uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems

Yuma Tsuta †, Naoki Yoshinaga ‡, Masashi Toyoda ‡
† University of Tokyo
‡ Institution of Industrial Science, University of Tokyo

Slide/code/dataset:
https://bit.ly/3hDwTj8
Open-domain dialogue systems become popular

Dialogue agents are used in daily life

• Dialogue agents are expected to reply to any user utterance (open-domain dialogues)

Cited from:
https://techcrunch.com/2019/02/05/report-smart-speaker-adoption-in-u-s-reaches-66m-units-with-amazon-leading/
How to develop open-domain dialogue systems?

• Large-scale human-human conversations on SNS helps to develop open-domain dialogue systems [Wu+2016]

• Automatic evaluation metrics are needed to develop dialogue systems efficiently
  • Existing reference-based evaluation metrics such as BLEU do not perform well on open-domain dialogue [Liu+2016]
Challenge in evaluating open-domain dialogue systems

Difficult to consider **all possible responses**
- **Diverse replies** can be allowed
- **Only one reference response** is available when real conversation data is used for evaluation

Input utterance

I watched movie yesterday

Reference response

Let me know how it was

Generated response 1

Let me know your impression

Generated response 2

What movie did you see?

• Evaluation with one reference response is unstable
Related work: ΔBLEU [Galley+2015]

ΔBLEU compute weighted BLEU using additional responses retrieved from Twitter and manual validation to those replies

Step 1. Retrieve reference responses from dialogue logs

Step 2. Rate reference responses by human annotator

Step 3. Compute weighted BLEU with human annotated test samples
STEP 1 on ΔBLEU: Retrieve dialogues as reference responses

Step 1. Retrieve reference responses from dialogue logs

Retrieve dialogues of which
- utterance is similar to input utterance (of test example), and
- response is similar to reference response (of test example)

based on BM25 as similarity function

Test example
- I watched movie yesterday
- Let me know how it was

Dialogue logs
- I went theater to see movie
- Let me know your impression

Step 2. Rate reference responses by human annotator

0.8
-0.6
0.3

Annotator
**STEP 2 on ΔBLEU:**
Rate reference responses by hand

**Step 1.** Retrieve reference responses from dialogue logs

**Dialogue logs (Twitter)**

**Test example**

**Query**

**Retrieved dialogues**

**Step 2.** Rate reference responses by human annotator

**Human annotator** rates retrieved responses in terms of appropriateness to a given utterance

**Utterance of test example**

I watched movie yesterday

**Retrieved responses**

- Let me know how it was
- Let me know your impression

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>-0.6</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
STEP 3 on ΔBLEU: Evaluate generated responses

Step 1. Retrieve reference responses from dialogue logs

Step 2. Rate reference responses by human annotator

Step 3. Compute weighted BLEU with human annotated test samples

Generated responses is
• rewarded when it matches positively rated reference response
• penalized when it matches negatively rated reference response
Issues on ΔBLEU

**Step 1. Retrieve reference responses from dialogue logs**

- Dialogue logs (Twitter)
- **Test example**
  - Query: 
    - User: \( U_1 \)
    - Response: \( R_1 \)
  - Retrieved dialogues:
    - User: \( U_2 \)
    - Response: \( R_2 \)

**Step 2. Rate reference responses by human annotator**

- Annotator:
  - Low semantical diversity of responses
  - Retrieval method based on the similarity of responses using BM25 (word overlap similarity function)

**Construction cost of test data**

- For the test of open-domain dialogue systems, test data on several domain is needed

**Test example**

- Query: 
  - User: I watched movie yesterday
  - Let me know how it was
- Response:
  - What movie did you see?
  - I went theater to see movie

**Dialogue logs (unlikely to be retrieved)**

- User: I watched movie yesterday
- Response: Let me know how it was

Proposed method: Automatic evaluation method $\upsilon$BLEU

Proposed method $\upsilon$BLEU deals with the issues on $\Delta$BLEU by

- Collecting more diverse reference responses, and
- Rating reference responses automatically

---

Step 1. Retrieve diverse reference responses

Step 2. Rate responses automatically by neural network (NN)-rater

Step 3. Compute weighted BLEU with automatically rated test samples
**STEP 1 on uBLEU:** Diverse responses retrieval

**Step 1.** Retrieve diverse reference responses

- **Query:** U₁
- **Dialogue logs (Twitter):** 
  - Retrieved dialogues: U₂, R₂
- **Test example:** I watched movie yesterday
  - Let me know how it was

**Step 2.** Rate reference responses by neural network (NN)-rater

- **Dialogue logs:** 
  - Retrieved dialogues: U₁, R₁
  - 0.8
  - -0.6
  - 0.3
  - NN-rater

**To semantically diversify retrieved responses,**
retrieve dialogues based on **similarity of utterances only**
  - **Cosine similarity** of averaged Glove [Pennington+2014] vector

**Test example:**
- I watched movie yesterday
- Let me know how it was

**Dialogue logs:**
- I went theater to see movie
- What movie did you see?
STEP 2 on υBLEU: Automatic response rating

Step 1. Retrieve **diverse** reference responses

- Dialogue logs (Twitter)
- Retrieve diverse reference responses
- Compute ΔBLEU with automatically rated test samples
- Rate reference responses by neural network (NN)-rater

Step 2. Rate reference responses by neural network (NN)-rater

- NN-rater classifies whether the conversation is appropriate
  - The utterance of text example and the response of retrieved dialogue
- Rate retrieved responses by the classification probability

**Test example**

- Query: U₁
- Dialogue logs (Twitter)
- Retrieved dialogues: R₁, U₂
- Retrieved dialogue: R₂

**Test example**

- Input: U₁, R₁
- Output: NN-rater (Bi-GRU+FFNN)
- Pₚₒₛ & Pₑᵣₑₙɡ
**STEP 2 on uBLEU:**
Training data of NN-rater

**Step 1.** Retrieve *diverse* reference responses

- Test example
- Query
- Dialogue logs (Twitter)
- Retrieved dialogues

**Step 2.** Rate reference responses by neural network (NN)-rater

- Positive sample: utterance which has several responses
- Negative sample: randomly sampled two conversations

- NN-rater is trained with automatically collected data
- Positive sample:
- Negative sample:

**Positive sample**

- Test example: U₁
- Query: U₁
- Retrieved dialogues: R₁, R₂

**Negative sample**

- Test example: U₂
- Query: U₂
- Retrieved dialogues: R₁, R₂
Experiment 1: Comparison of response retrieval methods

Evaluate the impact of collecting additional responses using the similarity of utterances

1. Compute **BLEU with reference responses** retrieved by changing the **target** and **function** to compute similarity
   - **Target**
     - Utterance & Response
     - Utterance only (proposal)
   - **Function**
     - BM25
     - Cosine similarity for averaged Glove vector (proposal)

2. Compare the correlations with human judgment and BLEU using each retrieved multiple reference responses
Experiment 1: settings

• Number of retrieved reference response: 15 responses

• Dialogue systems:
  VHRED [Serban+2017], C-BM25 (derivation of C-TFIDF [Lowe+2015]), human response

• Test data:
  100 pairs of Japanese conversations on Twitter in 2019

• Human annotation:
  • Five annotator rated 300 responses in terms of appropriateness in the scale of [1, 5]
  • Calculate Pearson correlation between individual judgment and each evaluation metric
  • Show the maximum and minimum value in five correlations
Experiment 1: Results

Pearson correlation between human judgment and BLEU with multiple reference response

- Max. / Min. of five correlations with individual judgements

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Pearson correlation</th>
<th>Metrics</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target</strong></td>
<td><strong>Function</strong></td>
<td><strong>Max.</strong></td>
<td><strong>Min.</strong></td>
</tr>
<tr>
<td>Original reference</td>
<td></td>
<td>0.276</td>
<td>0.190</td>
</tr>
<tr>
<td>Utterance &amp; Resp.</td>
<td>BM25</td>
<td>0.298</td>
<td>0.173</td>
</tr>
<tr>
<td><strong>Utterance only</strong></td>
<td>BM25</td>
<td>0.296</td>
<td>0.178</td>
</tr>
<tr>
<td>Utterance &amp; Resp.</td>
<td>Cosine sim.</td>
<td>0.322</td>
<td>0.177</td>
</tr>
<tr>
<td><strong>Utterance only</strong></td>
<td>Cosine sim.</td>
<td><strong>0.366</strong></td>
<td><strong>0.209</strong></td>
</tr>
</tbody>
</table>

Proposed method retrieves more beneficial reference responses than the method of ΔBLEU [Galley+2015]
Experiment 2: Comparison of evaluation metrics

Compare the correlations between human judgment and each evaluation metric

• Methods to compare
  • ΔBLEU [Galley+2015]
  • RUBER [Tao+2018]
  • υBLEU
  • RUBER with υBLEU (*)

* RUBER is constituted by
  • referenced-based metric and
  • unreferenced-based metric

we also propose automatic integrated metric (RUBER with υBLEU)
replacing its referenced-based-metric with υBLEU
Experiment 2: settings

• Training data of NN-rater and RUBER
  5.6M pairs of Japanese conversations on Twitter in 2017

• Dialogue systems, Test data and Human annotation
  Same as experiment 1
Experiment 2: Results

Pearson correlation between human judgment and evaluation metric

- Max. / Min. of five correlations with individual judgements

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max.</td>
</tr>
<tr>
<td>ΔBLEU</td>
<td>0.360</td>
</tr>
<tr>
<td>uBLEU</td>
<td>0.394</td>
</tr>
<tr>
<td>RUBER</td>
<td>0.325</td>
</tr>
<tr>
<td>RUBER with uBLEU</td>
<td>0.450</td>
</tr>
</tbody>
</table>

- Our proposal uBLEU outperformed ΔBLEU
- Integrating uBLEU into RUBER improved correlation
Summary

• Proposal:
  Uncertainty-Aware Automatic Evaluation Method for evaluating open-domain dialogue systems
  • Collect semantically diverse reference responses
  • Rate responses automatically with neural network classifier

• Using Twitter dialogues, we experimentally confirmed
  • Comparable with semi-automatic evaluation metric, ΔBLEU
  • Improve the correlation of the state-of-the-art automatic evaluation method RUBER by integrating with uBLEU
Acknowledgments

The research was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation