Exploring Question-Specific Rewards for Generating Deep Questions

Yuxi Xie¹ Liangming Pan²,⁴ Dongzhe Wang³
Min-Yen Kan² Yansong Feng¹,⁵

¹Wangxuan Institute of Computer Technology, Peking University
²School of Computing, National University of Singapore, Singapore
³Zhuiyi Technology
⁴NUS Graduate School for Integrative Sciences and Engineering
⁵The MOE Key Laboratory of Computational Linguistics, Peking University
Introduction

Challenges in Question Generation

- **Exposure Bias**
  - *Inconsistency* between training objectives and targets
  - Hard to measure the *global quality* of the generated questions
Introduction

Challenges in Question Generation

- **Evaluation**
  - Current n-gram based evaluation metrics cannot properly evaluate a question
  - Problems in Fluency, Relevance and Answerability still remain unsolved

Lawrence Ferlinghetti is an American poet, he wrote a short story named what?

Lawrence Ferlinghetti is an American poet, he is a short story written by who?

What mine was operated at an earlier date, Kemess Mine or Colomac Mine?

Between Kemess Mine and Colomac Mine, which mine was operated earlier?
**Introduction**

Reinforcement Learning in Question Generation

- **Decouple** the training procedure from the ground truth data
  - the space of possible questions can be better explored
- **Allow to target on specific properties** we want the question to exhibit during training
  - e.g. relevant to a specific topic or answerable by the document

- **How to define robust and effective QG-specific rewards** requires further investigation
  - optimizing the reward scores does not always lead to higher question quality in practice
Methodology

Three Research Questions to Answer

• Does optimizing RL rewards really improve the question quality from the human standard
• Which reward is more effective in improving the question quality
• How the rewards interfere with each other when jointly optimized
Methodology

Relevance Discriminator

• **Discriminator initialization**
  • BERT-based Sentence Classifier

• **Training datasets**
  • Positive: G.T. document + question
  • Negative
    • Basic: *question* swap
    • Ghost entity: *entity* swap between different samples
    • Logic correctness: *entity* swap within the same sample
Experiments

Automatic Evaluation

- Optimizing a single reward alone (F, R, A) improves the BLEU score and its corresponding reward score.
- The three rewards are correlated. One improves, the other two also increase.
- Jointly training multiple rewards in general leads to better performance.
- The increase in rewards do not correlate well with improvement on automatic metrics.

<table>
<thead>
<tr>
<th>Model</th>
<th>Rewards</th>
<th>BLEU1</th>
<th>BLEU4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>R-FLU</th>
<th>R-REL</th>
<th>R-ANS</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1. Baseline</td>
<td></td>
<td>33.68</td>
<td>13.46</td>
<td>21.39</td>
<td>35.06</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S1. F</td>
<td>✓</td>
<td>37.59*</td>
<td>15.22*</td>
<td>19.49*</td>
<td>35.08*</td>
<td>+1.48*</td>
<td>+0.49*</td>
<td>+0.03</td>
</tr>
<tr>
<td>S2. R</td>
<td>✓</td>
<td>36.33*</td>
<td>14.83*</td>
<td>20.63*</td>
<td>35.58*</td>
<td>+1.06*</td>
<td>+0.61*</td>
<td>+0.04</td>
</tr>
<tr>
<td>S3. A</td>
<td>✓</td>
<td>36.40*</td>
<td>13.95*</td>
<td>18.73*</td>
<td>34.07*</td>
<td>+1.30</td>
<td>+0.18*</td>
<td>+0.21*</td>
</tr>
<tr>
<td>E1. F + R</td>
<td>✓ ✓</td>
<td>37.82*</td>
<td>15.30*</td>
<td>19.95*</td>
<td>35.48*</td>
<td>+1.30*</td>
<td>+0.60*</td>
<td>+0.03</td>
</tr>
<tr>
<td>E2. R + A</td>
<td>✓ ✓</td>
<td>35.77*</td>
<td>14.46*</td>
<td>20.53*</td>
<td>35.26</td>
<td>+0.78</td>
<td>+0.49*</td>
<td>+0.36*</td>
</tr>
<tr>
<td>E3. F + A</td>
<td>✓ ✓</td>
<td>38.30*</td>
<td>14.99*</td>
<td>18.02*</td>
<td>34.50*</td>
<td>+1.71*</td>
<td>+0.40*</td>
<td>+0.51*</td>
</tr>
<tr>
<td>E4. F + R + A</td>
<td>✓ ✓ ✓</td>
<td>37.97*</td>
<td>15.41*</td>
<td>19.61*</td>
<td>35.12</td>
<td>+1.57*</td>
<td>+0.61*</td>
<td>+0.49*</td>
</tr>
</tbody>
</table>

Table 1: The QG performance evaluated by automatic metrics when separately or jointly optimizing for various rewards. The last three columns show the change of reward scores compared with B1, where R-FLU is the fluency, R-REL the relevance, and R-ANS the answerability rewards. * denotes that the increase/drop in performance compared with B1 is statistically significant for $p < 0.01$. 
**Experiments**

**Automatic Evaluation**

- If judging by *automatic evaluation metrics*, we find that optimizing QG-specific rewards is effective in generating deep questions, compared with other strategies.
- However, does optimizing rewards really improve the question **quality** as expected?

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>AE</th>
<th>LF</th>
<th>CP</th>
<th>CV</th>
<th>SA</th>
<th>RL</th>
<th>BLEU1</th>
<th>BLEU4</th>
<th>Meteor</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3. NQG++ (Zhou et al., 2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35.31</td>
<td>11.50</td>
<td>16.96</td>
<td>32.01</td>
</tr>
<tr>
<td>B4. Zhao et al. (2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35.36</td>
<td>11.85</td>
<td>17.63</td>
<td>33.02</td>
</tr>
<tr>
<td>B5. Zhao et al. (2018) + ans, cov</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38.74</td>
<td>13.48</td>
<td>18.39</td>
<td>34.51</td>
</tr>
<tr>
<td>B6. CGC-QG (Liu et al., 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31.18</td>
<td>14.36</td>
<td>25.20</td>
<td>40.94</td>
</tr>
<tr>
<td>B7. SG-DQG (Pan et al., 2020)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>40.55</strong></td>
<td><strong>15.53</strong></td>
<td>20.15</td>
<td>36.94</td>
</tr>
<tr>
<td>E4. Ours (F + R + A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37.97</td>
<td>15.41</td>
<td>19.61</td>
<td>35.12</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison. For all baselines, we use the reported performance from Pan et al. (2020). Legend: **AE**: answer encoding, **LF**: linguistic features, **CP**: copying mechanism, **CV**: coverage mechanism, **SA**: gated self-attention, **RL**: reinforcement learning.
Experiments

Human Evaluation

- Human ratings do not correlate well with automatic evaluation metrics
- Optimizing the relevance reward (S2) alone leads to an improvement of the human ratings for fluency, relevance, and answerability.
- Optimizing for answerability (S3) has a negative effect.

**Conclusion**: If we want to know whether a certain reward has an effect or not, judging from automatic metrics maybe deceiving.

**BUT**, why relevance works, but answerability fails?

<table>
<thead>
<tr>
<th>Model</th>
<th>Flu. (1-5)</th>
<th>Rel. (1-3)</th>
<th>Ans. (0-1)</th>
<th>Cpx. (1-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1. Baseline</td>
<td>3.98</td>
<td>2.77</td>
<td>0.67</td>
<td>1.59</td>
</tr>
<tr>
<td>S1. F</td>
<td>4.07</td>
<td>2.78</td>
<td>0.61</td>
<td>1.50</td>
</tr>
<tr>
<td>S2. R</td>
<td>4.24</td>
<td>2.83</td>
<td>0.70</td>
<td>1.51</td>
</tr>
<tr>
<td>S3. A</td>
<td>3.82</td>
<td>2.63</td>
<td>0.46</td>
<td>1.55</td>
</tr>
<tr>
<td>E4. F+R+A</td>
<td>4.10</td>
<td>2.72</td>
<td>0.53</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Table 3: Human evaluation results for different methods. Flu., Rel., Ans., and Cpx. denote the Fluency, Relevance, Answerability, and Complexity, respectively.
Experiments

Consistency between Rewards & Human Judgement

- the relevance rating has strong correlations with both the fluency rating and the answerability rating, compared with a relatively weak correlation exists between the fluency and answerability.

- the relevance reward has strong correlations with all three ratings
- the answerability reward has poor correlation with fluency and relevance.
Experiments

Consistency between Rewards & Human Judgement

- Relevance Reward: good correlation
- Fluency Reward: normal correlation
- Answerability Reward: bad correlation

Conclusion: how well the reward score correlates with the human judgement is a good way to know whether a certain reward works or not.
Experiments

Meso Analysis - Fluency

- sometimes the fluency reward is consistent with the human judgement on fluency
- the LM tends to assign low rewards to the questions with rare or unseen entities
- the lack of commonsense knowledge is another problem of the LM
Experiments

Meso Analysis - Relevance

- **two aspects** for the **relevance discriminator**
  - **ghost entity**
  - **logical inconsistency**
- **it is difficult** for the **model to assign a proper relevance score** when the question is asking about an **unmentioned aspect** of something in the document
  - **potential solution**: a good **answerability discriminator**
Experiments

Meso Analysis - Answerability

- most of the questions with high rewards are asking what year (the text highlighted in pink)
- when the question requires the QA model to conduct reasoning such as comparison and to utilize world knowledge, the QA model tends to give a low answerability reward
- to improve the answerability via a QA-based reward, it is crucial to address the QA model’s bias in prediction and improve its reasoning ability
Q & A

THANK YOU FOR WATCHING