Learning to Ask: Neural Question Generation for Reading Comprehension
Xinya Du, Junru Shao, Claire Cardie

What’s QG and why QG?

A New Task: Automatic natural question generation for sentences from text passages in reading comprehension.
Example: From Wikipedia article Oxygen

Sentence: Oxygen is used in cellular respiration and released by photosynthesis, which uses the energy of sunlight to produce oxygen from water.

Real Applications:

Education: Generating questions for testing understanding

Chat bot: asking questions to start a conversation or to request feedback.

Improving question answering (QA)

Sentence- and Paragraph-level seq2seq model

Task Objective:

• Input: a natural language sentence $x_t$, and optionally a natural language paragraph $x_p$.
• Objective: To generate a question about the input sentence, such that:

  1. We model the conditional next-word probability as:

     \[ P(y_t|x_t, y_{<t}) = \text{softmax}(W_s \cdot s_t + b_s) \]

     where $s_t$ is the context vector, sum of the weighted avg. of encoder hidden units.

   2. Encoding only sentence as input, do not consider paragraph/context-level information.

   3. Training: Minimize the negative log-likelihood with respect to $\theta$.

   4. Inference: Beam search and UNK replacement

   5. Sentence encoder (Section 4.1):

   \[ \text{Sentence encoder}(x_t) = \text{softmax}(W_s \cdot s_t + b_s) \]

   6. Paragraph-level model:

   • Encoding both sentence and paragraph (that contains the sentence) as input, but only attending source sentence hidden states.

   \[ \text{Paragraph encoder}(x_t, x_p) = \text{softmax}(W_s \cdot s_t + b_s) \]

   7. Output Analysis

   a. Green highlight shows the input sentence, which is used as input to both sent. and para. level models.

   b. Rule-based model copy does nearly word for word the input sentence with minor syntactic changes.

   c. Redundant info. sometimes ungrammatical

Interpretability

Attention weight matrix shows the soft alignment between the sentence (left) and the generated question (top).

In this example, for the decoded token, the input sentence token with highest attention is “elidon square shop”.

Media Coverage

• New Scientist: “Inquisitive bot asks questions to test your understanding”

• Tech Republic: “How researchers trained one AI system to start asking its own questions”

Conclusion

• We first proposed the first fully data-driven neural network approach for question generation in the reading comprehension setting. We investigated encoding sentence- and paragraph-level information for this task.

• Follow-up Work: Our EMNLP paper on sentence selection for passage-level QG, see you soon in Copenhagen :)!

• We release the processed dataset based on SQuAD.

Open question: How to better utilize QG for QA?

Dataset

We build the QG dataset based on the SQuAD corpus

Automatic Evaluation:

| Model               | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DirectIn</td>
<td>7.73</td>
<td>15.09</td>
<td>7.87</td>
<td>15.56</td>
<td>14.95</td>
<td>22.47</td>
</tr>
<tr>
<td>H&amp;S (rule-based)</td>
<td>3.13</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>29.75</td>
</tr>
<tr>
<td>H&amp;S (code-based)</td>
<td>13.61</td>
<td>3.64</td>
<td>1.00</td>
<td>0.38</td>
<td>10.47</td>
<td>17.62</td>
</tr>
<tr>
<td>Our model (pre-trained) + a paragraph</td>
<td>41.11</td>
<td>23.44</td>
<td>17.58</td>
<td>12.28</td>
<td>16.62</td>
<td>28.79</td>
</tr>
<tr>
<td>Ours (sent-level model)</td>
<td>50.75</td>
<td>30.44</td>
<td>26.48</td>
<td>17.89</td>
<td>27.95</td>
<td>39.37</td>
</tr>
<tr>
<td>Ours (para-level model)</td>
<td>51.25</td>
<td>34.14</td>
<td>28.75</td>
<td>19.05</td>
<td>29.75</td>
<td>40.89</td>
</tr>
</tbody>
</table>

Human Evaluation:

<table>
<thead>
<tr>
<th>Method</th>
<th>Natural</th>
<th>Difficulty</th>
<th>Avg. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>H&amp;S (rule-based)</td>
<td>1.94</td>
<td>3.04</td>
<td>3.04</td>
</tr>
<tr>
<td>Ours</td>
<td>1.86</td>
<td>3.04</td>
<td>3.04</td>
</tr>
<tr>
<td>Human</td>
<td>3.04</td>
<td>3.04</td>
<td>3.04</td>
</tr>
</tbody>
</table>

Two-tailed t-test statistical significance: *p < 0.05, **p < 0.01

Experiments

• Our sentence-level model beats the strong rule-based system.
• Directly copy (DirectIn) forms a very strong baseline.
• Pre-trained word embeddings help significantly.

• The neural model outperforms significantly rule-based methods by human evaluation.
• Larger margin compared with automatic, better automatic metrics to be designed.