Difficulty Controllable Question Generation: A New Task

Motivation:
- SQuAD questions have different difficulty levels. Q₁ is easy, Q₂ is hard.
- Can we control the difficulty of generated questions?

Task Definition:
- Given a sentence, a text fragment (answer) in the sentence, and a difficulty level
- To generate a question that is asked about the fragment and satisfy the difficulty level

Applications:
- Balance the number of hard and easy questions for knowledge testing
- Test how a QA system works for questions with diverse difficulty levels
- Improve performance of QA systems

Challenges
- No existing QA dataset has difficulty labels for questions
- For a single sentence and answer pair, we want to generate questions with diverse difficulty levels, but SQuAD only has one given question for each sentence and answer pair
- No metric to evaluate the difficulty of questions

Data Preparation

Question Difficulty is a subjective notion and can be addressed in many ways:
- Some stories are inherently difficult to understand
- Questions can be difficult in different ways, such as syntax complexity, coreference resolution and elaboration

Our Method for Data Preparation:
- Focus on generate SQuAD-like questions with diverse difficulty levels
- Two difficulty levels: Easy and Hard
- Develop an automatic labelling protocol
- Study the correlation between automatically labelled difficulty with human difficulty

Automatic labelling protocol:
- Employ two reading comprehension systems, R-Net and BiDAF
- A question would be:
  - labelled with ‘Easy’ if both R-Net and BiDAF answer it correctly
  - labelled with ‘Hard’ if both systems fail to answer it
- The remaining questions are eliminated for suppressing the ambiguity
- 44723 easy questions, 31332 hard questions

Human Rating on 100 Easy & 100 Hard Questions:
- 1-3 scale, 3 for the most difficult
- Easy: 1.90 vs. Hard: 2.52

Exploring Proximity Hints:
- If a question has more hints that can help locate the answer fragment, it would be easier to answer
- The average distance of those nonstop question words that also appear in the input sentence to the answer fragment

- Question Word Proximity Hints
  - The distance of nonstop question words are much smaller than the sentence words
  - Learn a lookup table to map the distance into a position embedding: (p₁, p₂, ..., pₙ)

- Difficulty Level Proximity Hints
  - The distance for hard questions is significantly larger than that for easy questions
  - Explore the information of question difficulty levels
  - Easy: (p₁, p₂, ..., pₙ); Hard: (p₁, p₂, ..., pₙ)

Characteristic-rich Encoder:
- Concatenate word emb and position emb: x = [w; p]
- Bidirectional LSTMs encode the sequence

Global Difficulty Control:
- Use style variable to initialize the decoder state: u₀ = [h₀, d]

Decoder with Attention & Copy

Experiment Results

Automatic Evaluation:
- Employ reading comprehension systems to evaluate the difficulty of generated questions
- N-gram based similarity: BLEU(B), ROUGE-(L-R), METEOR(MET)

Difficulty of the Generated Questions: Question Quality:

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Easy Questions (R:No)</th>
<th>Hard Questions (R:No)</th>
<th>Difficulty</th>
<th>Easy Questions (R:Yes)</th>
<th>Hard Questions (R:Yes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>1.25</td>
<td>1.80</td>
<td>Hard</td>
<td>2.40</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Controlling Difficulty:

Human Evaluation:
- Fluency (F) [1,2,3]: grammatical correctness and fluency
- Difficulty (D) [1,2,3]: difficulty of generated questions
- Relevance (R) [0,1]: if the question is ask about the answer

Experiment Results:

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Easy Questions</th>
<th>Hard Questions</th>
<th>Difficulty</th>
<th>Easy Questions</th>
<th>Hard Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion:

The difficulty-controllable question generation framework can control the difficulty of generated questions and adapt to different applications.

Future Work:
- Further explore the impact of difficulty on question comprehension and answering performance.
- Study the effect of different difficulty levels on human performance and satisfaction.
- Investigate the scalability of the framework to larger datasets and more complex scenarios.