Interactive Open-Domain Story Generation
Seraphina Tarrant, Haining Feng, Nanyun Peng
Information Sciences Institute, University of Southern California

Motivation
- Can human-machine collaboration improve open-domain neural story generation?
- Can it improve specific story aspects, as well as overall quality?

Previous approaches to human-machine collaboration offer limited interaction. We design a system that enables human interaction at multiple stages of the process: story-planning, story-writing, diversity controls*, and model-selection.

Self-reported (subjective)
Subjects self-report on their engagement, satisfaction with their story, and perception of story quality.

Independent Ranking
Independent human judges are asked to rank all stories from 1-5 under eight experiment conditions for Overall Quality, Relevance, Creativity, and Causal-Temporal Coherence.

Sample Interaction

Figure 1: full-interaction capabilities, annotated with user actions from an example study. Interaction is iterative: a user can edit or regenerate any element at any time.

We conduct user studies for multiple interaction scenarios. We constrain experiments to 10 minutes, and explore full-interaction, story-only, storyline-only, and diversity-only variations.

*diversity controls are softmax temperatures, which control the unusualness of system generations.

Code and data available at:
https://github.com/seraphinatarrant/plan-write-revise

Web Interface
- Topic
- Configuration
- Storyline
- Stories

System

We adapt the Plan-and-Write system; a storyline planning to story generation pipeline (Yao et al 2019) to enable interaction at the story-planning stage. We include their Title-to-Story baseline (no planning stage) and create a new Plan-and-Revise system, which incorporates two discriminators for Relevance and Creativity, as in Holtzman et al. (2018).

Metrics

ROC Stories: 98,162 commonsense stories data split into 8:1:1 for training, dev and test sets. Storylines are extracted via RAKE (a keyword extraction algorithm) as in Yao et al (2019).

Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Overall</th>
<th>Creative</th>
<th>Relevant</th>
<th>C-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine</td>
<td>2.34</td>
<td>2.08</td>
<td>2.46</td>
<td>2.54</td>
</tr>
<tr>
<td>Diversity only</td>
<td>2.50</td>
<td>2.96</td>
<td>2.75</td>
<td>2.81</td>
</tr>
<tr>
<td>Storyline only</td>
<td>3.21</td>
<td>3.27</td>
<td>3.88</td>
<td>3.65</td>
</tr>
<tr>
<td>All</td>
<td>3.76*</td>
<td>4.04</td>
<td>3.96*</td>
<td>4.24</td>
</tr>
<tr>
<td>All + Creative</td>
<td>3.73</td>
<td>3.96*</td>
<td>3.98*</td>
<td>3.83</td>
</tr>
<tr>
<td>All + Relevant</td>
<td>3.53*</td>
<td>3.52</td>
<td>4.05</td>
<td>3.91*</td>
</tr>
<tr>
<td>All + C-T</td>
<td>3.62*</td>
<td>3.88*</td>
<td>4.00*</td>
<td>3.98*</td>
</tr>
</tbody>
</table>

Table 1: Results for all experiments. Best scores per metric are bolded, scores not significantly different ($\alpha = 0.1$, per Wilcoxon Signed-Rank Test) are starred. C-T stands for Causal-Temporal Coherence, the + experiments are the extensions where the user focuses on improving a particular quality.

- humans tasked with improving a specific story aspect are successful at doing so
- interaction at both planning and writing stages improves story quality 10-50% over the less interactive baselines.
- additional interaction increases user self-reported satisfaction.