**SummScreen: A Dataset for Abstractive Screenplay Summarization**

Mingda Chen¹, Zewei Chu², Sam Wiseman³, Kevin Gimpel¹

¹Toyota Technological Institute at Chicago  ²University of Chicago  ³Duke University

**SummScreen**

- An abstractive summarization dataset combining TV series transcripts and episode recaps.
- We collected data from fan-contributed websites.
- We divide SummScreen into two subsets, FD and TMS, according to data sources.

<table>
<thead>
<tr>
<th></th>
<th>FD</th>
<th>TMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>#shows</td>
<td>88</td>
<td>10</td>
</tr>
<tr>
<td>#episodes</td>
<td>4348</td>
<td>22503</td>
</tr>
<tr>
<td>#lines in trans.</td>
<td>447.6</td>
<td>360.8</td>
</tr>
<tr>
<td>#char. utterances in trans.</td>
<td>330.7</td>
<td>327.0</td>
</tr>
</tbody>
</table>

**Transcript from “The Big Bang Theory”:**

[H] The apartment

Sheldon: What color would you like to be?
Leonard: Well, I’d like to be green, but you know you always take it.
Sheldon: That’s not true. Any color’s fine with me. Yeah, I could be a - a combination of blue and yellow.
Leonard: Blue and yellow make green.
Sheldon: Well, then it’s settled.

Penny: Hi. Ready to go?
Sheldon: Oh, good news, we ordered lunch, so we can all stay here and play Lord of the Rings Risk.
Amy: Sheldon, we said that we would play games with you tonight.
Sheldon: Oh, no, we’ll still be playing it tonight, this game can easily take eight hours.
Penny: Sweetie, you really thought I’d want to do this?

Leonard: No.
Penny: Well, did you tell him that?
Leonard: Yes.
Penny: Did you say it out loud with words?
Leonard: No.
Penny: I do n’t want to spend the whole day playing a board game.

**Human-Written Recap:**

Sheldon and Leonard are happy playing a board game until Amy and Penny say they are tired of doing what the guys want ...

**Transcript from “Doctor Who”:**

**Line 119** DOCTOR: Camera! Camera! (takes camera from ALEC ’s unresisting hands)

**Line 212** The DOCTOR turns around and continues to take photos with the camera …

**Line 287** The DOCTOR steps out of the TARDIS wearing the spacesuit … He scans with the sonic before picking up the camera …

**Line 336** DOCTOR: Right! Done! That’s it … She’s a pioneer, a time traveller - or at least she will be, in a few hundred years.

**Human-Written Recap:**

The Doctor borrows Alec’s camera and uses the TARDIS to take pictures of the mansion’s location throughout time. Thanks to this, the Doctor learns it’s not a ghost in the pictures, but a time traveler named Hila Tacorian …

**Results**

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNM</td>
<td>7.9</td>
<td>31.3</td>
<td>7.8</td>
<td>27.4</td>
<td>18.6</td>
</tr>
<tr>
<td>Neural Model</td>
<td>2.6</td>
<td>25.9</td>
<td>4.2</td>
<td>23.8</td>
<td>14.1</td>
</tr>
<tr>
<td>Hybrid Model</td>
<td>2.4</td>
<td>25.3</td>
<td>3.9</td>
<td>23.1</td>
<td>13.7</td>
</tr>
<tr>
<td>Hybrid Model (w/ oracle content selector)</td>
<td>3.0</td>
<td>26.4</td>
<td>5.0</td>
<td>23.3</td>
<td>14.4</td>
</tr>
</tbody>
</table>

**FD Test Results**

<table>
<thead>
<tr>
<th></th>
<th>BoC-p</th>
<th>BoC-r</th>
<th>BoR-p</th>
<th>BoR-r</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNM</td>
<td>56.7</td>
<td>59.2</td>
<td>28.2</td>
<td>29.4</td>
<td>43.3</td>
</tr>
<tr>
<td>Neural Model</td>
<td>54.7</td>
<td>38.5</td>
<td>22.8</td>
<td>15.1</td>
<td>32.8</td>
</tr>
<tr>
<td>Hybrid Model</td>
<td>61.2</td>
<td>51.4</td>
<td>29.8</td>
<td>23.6</td>
<td>41.5</td>
</tr>
<tr>
<td>Hybrid Model (w/ oracle content selector)</td>
<td>70.0</td>
<td>57.8</td>
<td>36.9</td>
<td>29.1</td>
<td>48.5</td>
</tr>
</tbody>
</table>

**Approaches**

- Neural models
- Nearest neighbor models (NNM): retrieve summaries from the training set.
- Hybrid models (nearest neighbor content selector → neural models)

**Metrics**

- BLEU and ROUGE
- Entity metrics:
  - Bag of characters (BoC): the fraction of the characters overlapping with gold
  - Bag of character relations (BoR): the fraction of the cooccurred character pairs overlapping with gold

**Dataset Challenges**

- Understand the context in which characters’ utterances are situated.
- Draw information from a wide range of the input transcripts.