Is It Fun and Fair?

Evaluating Multimodal Storytelling Robot in Dementia-Care

Author:

Konstantin Teplykh
k.teplykh@student.tudelft.nl

Supervisors:

Mark Neerincx
Paul Raingeard de la Bletiere

Figure 1: Storytelling Robot

I. Aim

To assess whether a storytelling robot:

- produces **outputs without data bias**
- **represents** participants **equally**
- provides an **enjoyable experience**

II. Nature of Data Bias

- **Semantic Bias**: the output misrepresents the mood or theme of the input.
- **Factual Consistency Bias**: specific details are incorrect or missing.

III. Storytelling Session

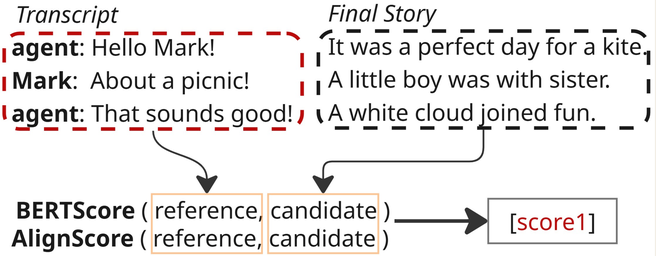
- Outputs: synthesized story, generated image and song, ground-truth conversation transcript.

IV. Proposed Evaluation Pipeline

Story Analysis

- Semantic consistency via **BERTScore**
- Factual consistency via **AlignScore**

Figure 2: Example Strategy



Enjoyment Analysis

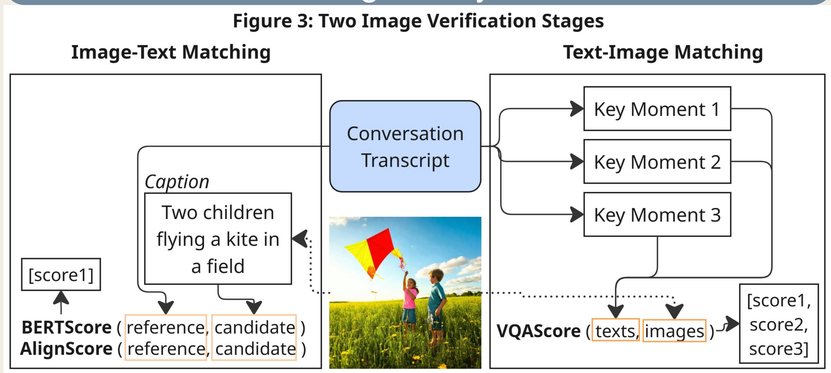
- Emotional engagement via language cues from conversation
- Utterance Emotion Dynamics framework

Audio Emotion Analysis

- Emotional coherence across outputs
- MTG valence-arousal model for song

Image Analysis

Figure 3: Two Image Verification Stages



V. Experiments

- Experiment 1: **Coherent Generation**. (Mark, Jen, robot)
- Experiment 2: **Biased Generation**. Same session setup, but outputs altered.

VI. Results

Figure 4: Story Analysis.
BERTScore and AlignScore across both experiments (Aggregated Transcript vs. Full Story strategy).

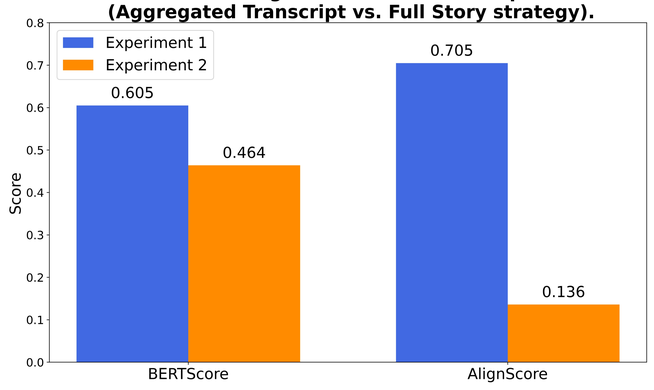


Figure 5: Story Analysis.
Per-speaker BERTScore and AlignScore in both experiments (Per-Speaker Utterance vs. Full Story strategy).

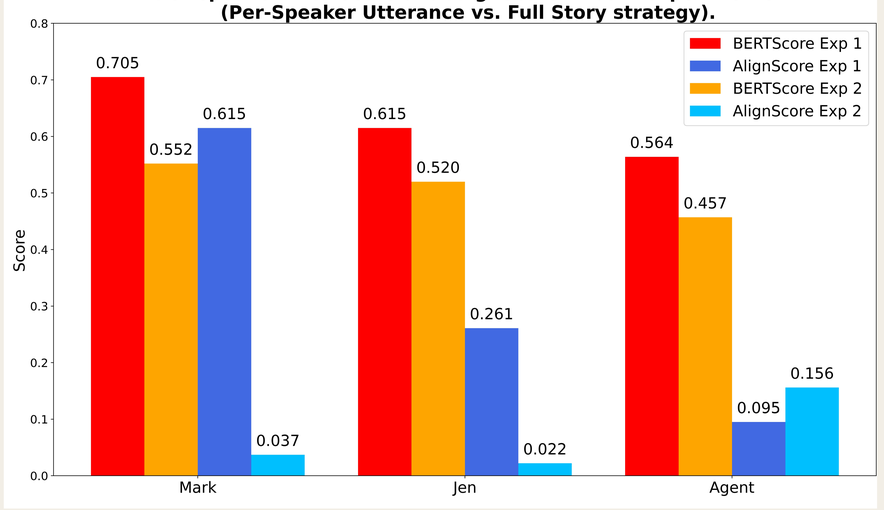
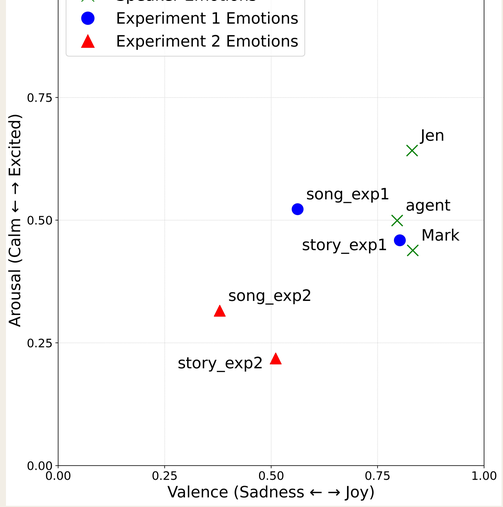


Figure 6: Mood Map - Energy vs. Positivity



VII. Conclusion

- Pipeline provides clear, scalable method for evaluating storytelling robots.
- Approach lays the groundwork for future research on fair and meaningful content generation in dementia care.

TU Delft