Correspondence between perplexity and human evaluation of generated TV-Show scripts.

Pia Keukeleire, Tom Viering, Stavros Makrodimitris, Arman Naseri Jahfari, Marco Loog, David Tax
TU Delft, P.Keukeleire@student.tudelft.nl,{T.J.Viering, S.Makrodimitris, A.Naserijahfari, M.Loog,D.M.J.Tax}@tudelft.nl

1 Background

Natural language generation:
- Chatbots
- Automatic image description

TV-show script generation:
- Inspiration for writers
- Challenge computers abilities to:
  - Mimic human creativity
  - Achieve global coherence

2 Questions

How to evaluate a generated scipt?
- Human evaluation = time consuming
- Perplexity (PPL): a measurement of how well a probability distribution or probability model predicts a sample.

How does perplexity correspond to human evaluation?

3 Methods

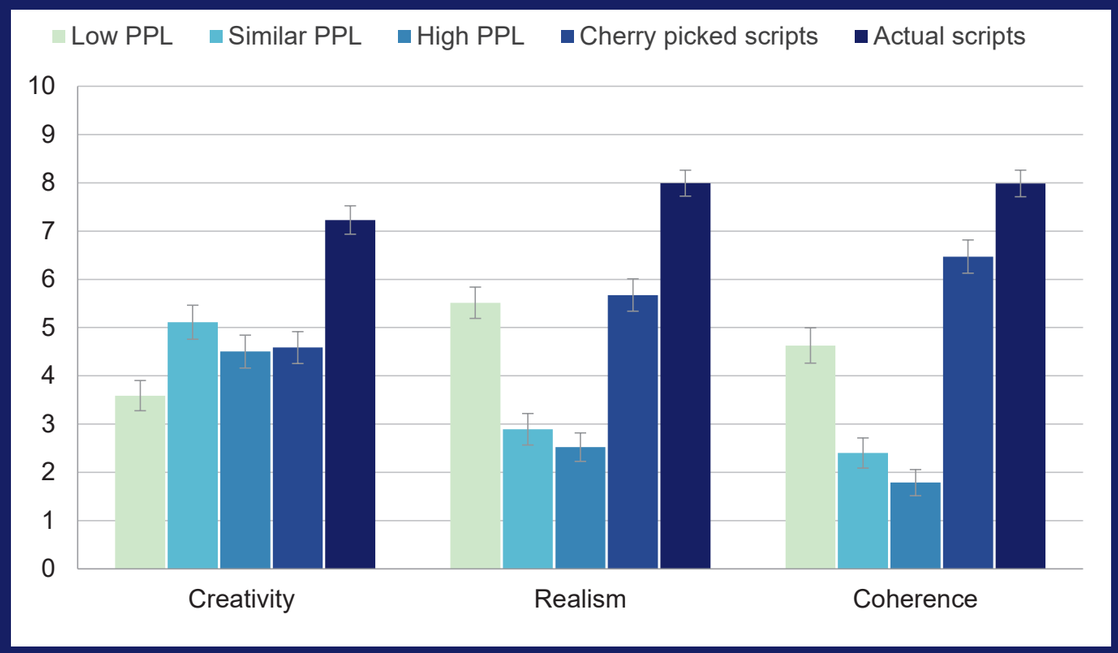
- OpenAI's GPT-2 model
- Finetuned on *Friends* scripts
- 81 different generations
- Survey which included:
  Scripts with:
  - low PPL value
  - PPL value similar to test set
  - High PPL value

  two cherry picked scripts
  two actual scripts.
- Assessment on:
  Creativity, realism and coherence

Example of Cherry Picked Generated *Friends* Script

```
Phoebe: Okay, so, uh, so how's your body?
Joey:   Oh, it's great! I'm feeling great!
Phoebe: Okay, so how about your teeth?
Joey:   Oh, you know, they're like a little harder.
Phoebe: Yeah, well, you can see my teeth now. (laughs)
Joey:   So, how's your eye?
Phoebe: Oh, it's great.
Joey:   So, how's your nose?
Phoebe: Oh, it's great. (laughs)
Joey:   So, how's your lip?
Phoebe: Oh, it's so soft.
```

Creativity, Realism and Coherence for each Type of Script



4 Results

Generations:
PPL between 1.13 and 7.85
Actual scripts:
PPL between 4.88 and 5.94

226 survey answers indicate that:
Low PPL value
= High Coherence and Realism score
Similar PPL value
= High creativity score

5 Discussion

- Not enough different PPL values researched to discover a trend
- Results may be different for other generation tasks
- Does not show if realistic Friends script generations exist
- Does not account for multiple possible participant biases

6 Conclusion

- Trade-off between creativity and realism/coherence is big
- Perplexity is not adequate, but interesting to measure variance in text
- Perplexity should be combined with other metrics

TU Delft

CSE3000 · 25.06.2020