What Makes a Good Story:
Analyzing Aspects of Story Evaluation

Yana Shishkina

1 Introduction

Narrative generation is one of the most creative and challenging tasks in NLP. An enjoyable story is not just grammatically correct and interconnected, but it id a combination of several aspects. It is important to specify these aspects so that researchers who aspire to create engaging stories can focus on the specific factors that might be lacking in their work. This work suggests a number of factors that might influence the overall evaluation of a story and aims to discover which of them contribute more to the story's overall quality and which ones are correlated. The work also addresses whether human-written stories are better than those generated with currently available language models according to the chosen aspects.

2 Data

The data used in this study is based on human-written English stories from the ROCStories dataset (Mostafazadeh et al., 2016). These stories consist of five-sentence short texts that align strongly with common sense, as the main purpose of the corpus was to collect logical and coherent narratives. To extend the dataset for additional research, I included stories generated by a model fine-tuned on the ROCStories dataset. The complete dataset contains 1162 stories, including 511 written by humans and 651 generated ones. All of the stories were evaluated by English-speaking annotators on a number of fine-grained criteria using the crowdsource platform Toloka.

The annotators were asked to evaluate the stories based on overall quality, creativity, and coherency, which were adopted from a previous work on narrative generation (Goldfarb-Tarrant et al., 2019). Additionally, aspects such as fluency and alignment with common sense were also considered. More granular aspects of creativity, such as the presence of a plot twist and the likelihood of a person sharing the story with others, were also evaluated, as well as the presence of a narrative arc - whether the story had a beginning, culmination, and end. The full set of assessed factors is presented in Table 1. Furthermore, a text classifier was used to automatically determine which emotions were prevalent in each story and whether there were any common emotional arcs represented by changes in emotion.

Table 1: Human evaluation criteria and their scales

Criterion	**Scale**
Overall quality Fluency Interest score Coherency	[0-4] (terrible – great)
Share score	[0-2] (yes/no/not sure)
Common sense Narrative arc Plot twist Human-like	[0-1] (yes/no)

3 Hypotheses

The main hypothesis is that all aspects contribute to the overall quality of the stories, but to different extents. It is expected that there will be a positive correlation between the criteria and the quality.

It is also expected that some aspects will be positively correlated with others, such as plot twists with interestingness, and human-likeness with fluency.

When comparing human-written and generated stories, the hypothesis is that human-created texts will be significantly better, but that the narratives in the ROCStories dataset may also not be perfect in some aspects.

Emotional components of the stories will also be analyzed, and the hypothesis is that the specific emotional arcs named Shapes of the stories by Kurt Vonnegut will contribute more to the success of a story among humans.

References

Goldfarb-Tarrant, S., Feng, H., and Peng, N. (2019). Plan, write, and revise: an interactive system for open-domain story generation. *arXiv preprint arXiv:1904.02357*.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.