Evaluation of Interest and Coherence in Machine Generated Stories

Dominic Callan and Jennifer Foster

School of Computing, Dublin City University, Ireland
dominic.callan24@mail.dcu.ie, jennifer.foster@dcu.ie

Abstract. Evaluation of the narrative text generated by machines has traditionally been a challenge, particularly when attempting to evaluate subjective elements such as interest or believability. Recent improvements in narrative machine text generation have been largely driven by the emergence of transformer-based language models, trained on massive quantities of data. In this study, a corpus of stories is generated using the pre-trained GPT-Neo transformer model, with human-written prompts. The stories generated through this process are subsequently evaluated through both human evaluation and two automated metrics: BERTScore and BERT Next-Sentence-Prediction. The results show variation in human evaluation results in comparison to automated metrics, suggesting further work is required to train automated metrics to identify text that is defined as interesting by humans.

Keywords: NLP · NLG · Machine-Generated Text · Transformer · Evaluation

1 Introduction

Many challenges exist in the evaluation of machine generated text. With the improvement in text generation quality by modern transformer models [22], fluency has greatly increased, however evaluating other elements of the text through automated metrics continues to prove difficult. Story generation differs significantly from other machine text generation challenges; rather than focusing on word overlap with an input or reference text as the metric for success, developing a believable narrative text requires composing coherent natural language texts that describe a sensible sequence of events [23]. A 'good' or successfully generated story is a subjective idea; there are many criteria that should be considered, with the result that the evaluation of stories is a difficult problem that is relatively understudied [15].

In other text generation tasks, such as Machine Translation, 'gold-standard' reference texts exist as a benchmark for comparison. No equivalent baseline reference texts exist against which to compare machine-generated stories when evaluating subjective concepts such as creativity or interestingness; creative language cannot easily be defined in this way, as evaluating in this manner does not allow for the possibility of correct but novel generation [19]. Clark et al. [6]

attempted a comparison of human authored and machine generated text across several domains, including story and news generation. They observe that human evaluators focused on form and structure rather than content in deciding whether a text was written by a machine or a human. This allowed for the conclusion that machines write fluently but did not address other narrative strengths of the text produced.

As noted by Akoury et al. [1] and Roemmele et al. [20], a potentially infinite number of human-generated stories can be generated that have attributes that may be considered interesting by human evaluators. To use a small sample of these as a reference for evaluating interest is therefore unreliable; Attempting to base a model against one human-generated reference could bias the model towards a certain style of writing or type of vocabulary, when the goal is to evaluate how interesting the story is.

This paper takes the dual approach of obtaining human judgements for a set of machine-generated stories, focusing on criteria around story interest, coupled with automated evaluations of the same texts. The automated metrics implemented focus on semantic similarity estimation, rather than n-gram overlap. A goal of the study is to evaluate the success of such automated metrics on narrative text generated by a large-scale transformer model and review these results in comparison to human evaluation of the same stories.

2 Background

2.1 Defining Story Interest

To evaluate the extent to which subjective attributes like interest, creativity or believability are applicable in machine generated text, certain criteria must be defined as metrics. In their in-depth study of human evaluations of automatically generated text, van der Lee et al. [14] reported that the most used metrics in these types of studies were fluency, naturalness, quality, and meaning-preservation, but ultimately, they note that the criteria chosen should depend on the specific task. Gatt & Krahmer [10] produce a similar list, also including 'Human-likeness' and 'Genre compatibility'. Celikyilmaz et al. [4] discuss certain criteria and attributes that should be present, including overall style, formality, or tone of the generated text. They add that there should be a 'typicality' to the generated text, meaning that it should be the type of text that we often see. Accuracy is of less concern for story-ending generation, as their output cannot usually be judged by fidelity to an identifiable, external input [13]. Grammaticality and fluency are not significant problems with modern transformer-based systems in comparison with older systems – the errors are instead often semantic or narrative [23]; humans can easily recognise non-sequitur sequences of events or conclusions, even when they are grammatical [19]. The difference between well written, coherent text, and interesting text is difficult to define. Generating text that simply describes a sequence of events alone is not enough for it to be considered interesting and coherent [16].

2.2 Human Evaluation of Machine Generated Text

NLG evaluation has long been identified as a difficult and complex area to measure accurately [12]. Human evaluation is still considered as the benchmark for evaluating machine generated outputs [11, 13]. Chaganty et al. [5] note that the many problems with automated evaluation metrics motivate the need for human evaluation. A goal of natural language generation is to produce fluent outputs that can be read by laypeople [23]; it is suitable therefore that this same group of 'laypeople' review the output where possible. We lack a good way of encoding aspects of what constitutes human quality output so we must rely on human evaluation of our models [6]. However, undertaking human evaluation of machine generated text systems also involves many challenges. Human interaction can be slow, is expensive and often hard to scale up [15, 5]; Purdy et al. [19] observe that the cost of human evaluation presents a bottleneck to AI research on story generation. Training of evaluators on what to expect and setting context and expectations can help them to focus on specific features of the text, which can be necessary given a tendency for humans to focus on form and fluency ahead of content [6].

2.3 Crowdsourcing

When crowdsourcing human evaluations, Celikyilmaz et al. [4] highlight issues with using sources like Amazon Mechanical Turk, especially when the task is to evaluate longer text sequences. These workers are typically more used to evaluating microtasks and may be less experienced with evaluating stories. Strong clear guidelines and instructions need to be issued to maximise the effectiveness of these evaluations. Lowe et al. [15] however warn that there must be a balance, as too much instruction can introduce bias. Van der Lee et al. [13] caution that there is a risk of inadvertently recruiting bots or participants who want to get paid for as little work as possible.

2.4 Automatic Evaluation of Machine Generated Text

The many challenges around reliable and scalable human evaluation have driven the development of automated evaluation systems. However, this challenge has traditionally proven difficult in NLG; text generation can go wrong in different ways while still receiving the same scores on automated metrics [13]. Many automated metrics exist currently. BLEU [17] has traditionally been used in NLG systems to evaluate word overlap, however it is not a suitable metric for measuring the success of developing narrative text. Chaganty et al. [5] note that while BLEU is cheap to run, it correlates poorly with human judgement. By rewarding word overlap, BLEU assigns a positive value to repetition, an element of machine text-generation that is to be avoided in story generation. As a metric, BLEU breaks down when the space of allowable outputs is large, as in open-ended generation like with prompts and stories [23]. Other metrics have emerged; BLEURT [21] is a BERT-based evaluation metric that is fine-tuned

on synthetically generated sentence pairs using automatic evaluation scores such as BLEU. It is then further fine-tuned on machine generated texts and human written references using human evaluation scores and automatic metrics as labels. The Automatic Dialogue Evaluation Model (ADEM) proposed by Lowe et al. [15] is a model-based evaluation that is learned from human judgements. It is mainly used for evaluating dialogue generation and is shown to correlate well with human judgement. Hashimoto et al. [11] propose Human Unified with Statistical Evaluation (HUSE), focussing on open ended text generation tasks, such as story generation. This model combines statistical evaluation and human evaluation metrics in a single model and differs to ADEM in this way.

3 Methodology

3.1 Input Data

The dataset used is a set of prompts taken from the reddit.com 'writing prompts' data set, introduced by Fan et al. [8]. The themes of the prompts vary, although they are often centred around fantasy or sci-fi. The average prompt length is 147 characters or 27 words. The shortest is 8 words and the longest is 56.

3.2 Transformer Language Models

Transformer language models make use of an 'attention' function which helps to identify for each word how relevant other words in the sequence are. The transformer architecture is used in the BERT system developed by Google [7], and in GPT-2 and GPT-3, developed by OpenAI [3]. For this study, the GPT-style architecture is implemented for the text generation process and BERT is used to underpin the automated evaluation of the machine generated text. Although both are Transformers, there are fundamental differences in how the two systems operate; BERT is trained to predict a masked token given the tokens on its left and right, and to predict whether two sequences follow on from each other. GPT is trained to predict the next token in a sequence, where every token can only attend to context to its left.

Licencing costs prevented the use of the GPT-3 model for this study. The GPT-Neo 2.7B parameter transformer model is used instead. Developed by EleutherAI, it is designed to be an open-source replication of Open-AI's GPT-3 architecture [2]. GPT-Neo model is trained on 'The Pile' dataset, an 825GB diverse open-source English text corpus targeted at training large scale data models [9]. The Pile is made up of 22 smaller datasets, including BookCorpus2, YouTube closed-captions, Project Gutenberg, and English Wikipedia. 800 stories were generated for this study. Given that the focus of this analysis is on narrative style text, when either the prompts or the stories were of a non-narrative nature, they were excluded from the final corpus. From the remaining corpus of narrative-style stories, 100 prompt-story pairs were chosen at random for evaluation. The average story length is 77 words, the longest has 96 words and the

shortest has 54. A cap of 400 characters was used and the average character count is 381 characters. This cap was implemented both as a method of maintaining coherence but also as a consideration to the survey participants who would be reviewing each story.

Automated Evaluation Metrics Whilst it is clearly identified in literature that good automatic evaluation metrics are still hard to come by ([13], [18],[19]), we chose the following two metrics for this study: BERTScore and BERT Next Sentence Prediction. Upon review of the available automated metrics, these were chosen since both focus on semantic similarity and thus have the potential to capture some notion of story coherence.

BERTScore is a language generation evaluation metric based on BERT [24]. It calculates a similarity score for two sentences, as a sum of cosine similarities between the contextualised word embeddings produced by a pretrained BERT model for each word in each sentence. By assigning different embeddings to words depending on their surrounding context, BERTScore attempts to reward semantic relationships between an input and an output, a core element of successful story generation. Example 1 shows a prompt/story sentence pair from the data that achieves a high re-scaled BERTScore of 0.287 and low Cumulative 1-gram BLEU score: 0.1226:

Example 1. ***Prompt:*** *You are born with the ability to stop time, but one day you see something else is moving when you have already stopped time.* ***Story sentence:*** *Your brain takes over and tells you to move, but you can't.*

Whilst it was developed for image captioning and machine translation tasks, BERTScore is designed to be task-agnostic. It is unclear how it can perform on open-ended tasks.

BERT Next Sentence Prediction BERT is trained on two tasks, BERT Masked LM and BERT for Next Sentence Prediction [7]. NSP is the task of predicting the probability that a sentence logically succeeds the previous sentence and is designed to learn the relationships between sentences. For this study, this BERT-NSP model is implemented as the second automated metric to evaluate stories. Each sentence pair is tokenised, and the BERT model processes the sentences and outputs 0 to indicate that Sentence-Two does follow Sentence-One, and 1 when it believes it does not.

3.3 Implementation of Evaluation Metrics

BERTScore The BERTScore metric tokenises two selections of text that are to be compared, and using contextual embeddings, derives a semantic similarity metric by calculating cosine similarities between the embeddings.[1] Two approaches were undertaken to obtain two BERTScore metrics for each story. In

[1] Zhang et al. [25] announced an optional improvement to BERTScore after the release of their original paper, to address the relatively small range observed between high

the first approach, the BERTScore is calculated between *each sentence in the story and the prompt*, and the resulting scores are averaged. This score is identified as BERTScore-1 in the results. The second approach compares *each sentence to the previous sentence*, rather than comparing each sentence to the prompt. These scores were again aggregated and are captured as BERTScore-2. By taking this approach, it can be observed firstly if individual segments of the story are semantically linked back to the prompt, but also if each segment is semantically linked to the previous segment.

BERT Next Sentence Prediction Similarly to BERTScore, BERT-NSP evaluates the prompt/story pairs in two different ways. BERT-NSP-1 looks at predicting whether each sentence in the story logically follows on from the prompt, whereas BERT-NSP-2 compares the prompt to the first sentence of the story, and then each subsequent sentence to the previous sentence.

Human Evaluation by Survey Human evaluation was undertaken using anonymous surveying where participants were firstly advised that the stories were written by machines. For each pair, the evaluators were shown the prompt and the subsequent story generated by the GPT-Neo model and asked to assess it on a Likert-scale with a score of between 1 and 7 for the following four questions:

1. How related do you think the story is to the prompt?
2. How much sense does the story make to you?
3. How interesting is the PROMPT to you?
4. How interesting is the STORY to you? (Would you read more?)

There was also a further optional free-text question at the end of each survey, for evaluators to leave general comments or impressions. The 100 prompt/story pairs were split into five sets of 20 pairs to reduce the chances of evaluators tiring or growing bored and abandoning the survey. The wording of these questions is designed to ask in plain-English terms about the coherence and interestingness of the stories generated by the machines. It was important to record the perceived semantic connection between the prompt and story; an interesting story could be produced by the system, however if it did not relate to the prompt then the objective of the task has not been achieved. A question on the story making sense to the evaluator was a proxy for story-coherence. This was introduced to observe whether a story needs to be coherent to be interesting to a reader, or conversely if an incoherent story was likely to be deemed uninteresting. Separate to the interest-level of the story, evaluators were asked if they found the prompt interesting, as their level of interest in the prompt may impact their interest in

and low scores. They suggest that the cosine similarity score is rescaled through a linear transformation, noting that this rescaling doesn't negatively impact correlation with human judgement. This rescaling is implemented in BERTScore calculations in this paper.

the resulting story generated. Two sample prompt/story pairs were included in the instructions of the survey, to provide context on the type of text that the evaluator would be reading in the survey and to set their expectations. Each prompt/story was reviewed by a minimum of 6 unique reviewers, although the majority were reviewed by 7 or more.

4 Results and Discussion

Table 1. Correlation between human judgements and automated metrics.

Metric	Q1	Q2	Q3	Q4	BS1	BS2	NSP1	NSP2
Q1	1.00	0.72	0.24	0.62	0.41	0.29	0.25	0.25
Q2	0.72	1.00	0.19	0.80	0.25	0.21	0.12	0.15
Q3	0.24	0.19	1.00	0.34	0.09	0.12	0.02	0.06
Q4	0.62	0.80	0.34	1.00	0.28	0.23	0.10	0.25
BS1	0.41	0.25	0.09	0.28	1.00	0.66	0.32	0.26
BS2	0.29	0.21	0.12	0.23	0.66	1.00	0.24	0.27
NSP1	0.25	0.12	0.02	0.10	0.32	0.24	1.00	0.61
NSP2	0.25	0.15	0.06	0.25	0.26	0.27	0.61	1.00

Table 1 shows a matrix illustrating correlation between the different survey questions and the four automated metrics implemented.

4.1 Human Evaluation

Within the human survey results, the strongest correlation of 0.80 is between story coherence (Q2) and story interest (Q4), suggesting that evaluators were most interested in the stories that they found to be the most coherent. A strong relationship (0.72) is also observed between the story coherence (Q2), and the story-prompt relationship (Q1), indicating potentially that evaluators factored in the connection between prompt and story when considering overall coherence; a story that was coherent, would be deemed less so if it did not follow on logically from the prompt. From observing average scores received for each question illustrated in Table 2, the prompt-interest received the highest average rating of 4.45 out of 7; the prompts were seen as more interesting than the stories that were generated. This suggests that, in general the model was unable to produce stories of greater interest than the prompts; it should be noted that there was a low positive correlation of 0.34 reported between prompt-interest and story-interest, showing little connection between these two.

There was no preference observed for either longer prompts or longer stories. The lowest average score for any question was for story interest (Q4), at 3.26/7. Focusing on average ratings across the four questions, the highest rated prompt/

story pair can be seen in Example 2. The story remains coherent and relevant to the prompt throughout, not deviating, even at its latter stages: both coherence and prompt-relationship are shown to correlate with story interest. Introducing the term 'corporate overlords' puts a somewhat dystopian slant to the narrative that wasn't necessarily part of the prompt, showing progression of the story in a defined and novel direction, which the evaluators seemed to have enjoyed. The lowest rated prompt/story pair, observed in Example 3, displayed few of the above-mentioned positive characteristics. The story does not relate back to the prompt and the context of the story contradicts the prompt semantically.

Table 2. Human Evaluation mean scores from survey results.

Question	Average Score
Q1. Story/prompt relationship	3.74/7
Q2. Story Coherence	3.42/7
Q3. Prompt Interest	4.45/7
Q4. Story Interest	3.26/7

Example 2. ***Prompt:*** *In the future, Disney purchases an entire planet and makes the whole thing into one giant amusement park.* ***Story:*** *The problem is that a large portion of the population hates the park, hates its corporate overlords, hates its theme park, and is starting to revolt. At the park itself, the people are divided. Some people are happy, some people are not. One thing that makes this story different from movies like the Matrix is that the people who live at the park have no control over the things that are happening there.*

Example 3. ***Prompt:*** *Everyone has disappeared but you, you think that everyone is just holding surprise party for you.* ***Story:*** *Then, the story starts. The film follows the adventures of jinx, a former pro wrestling manager who is now a private chef. His wife. jenna leigh, had just found out that he was dating a woman named dana bjorn Stresemann. A few months later, jinx and dana get back together after a break-up. Now, when jinx is going away for work, dana contacts the police with a missing person notice.*

4.2 BERT Next Sentence Prediction

The BERT-NSP scores are relatively high; in most cases both BERT-NSP-1, where each sentence relates back to the prompt, and BERT-NSP-2, where each sentence is compared to the previous sentence, predict that the next sentence does logically follow the preceding sentence. Within the 100 prompts/stories assessed, there were a total of 531 sentence pairs reviewed for next sentence prediction combinations. For BERT-NSP-1, 433 of the comparisons were deemed to be logical next sentences and only 98 were not. For BERT-NSP-2, an even

higher number of sentences were predicted to follow on from the previous one; 497 of the 531 sentences. This is somewhat expected, given the similarities in training objectives (predicting the next token versus next sentence prediction) and training data between GPT-Neo and BERT.

4.3 BERTScore

For both BERTScore metrics, cosine similarity was calculated for each sentence pair and then averaged for an overall score for a given prompt/story. The BERTScore-1 results ranged from -0.187 to 0.365 with a mean of 0.074, and the BERTScore-2 results ranged from -0.168 to 0.439 with a mean of 0.138. Some of the highest BERTScore results were for stories that demonstrated a notable amount of repetition, e.g Example 4 which received the highest BERTScore-1:

Example 4. ***Prompt:*** *A dozen small alien ships enter the solar system, they ignore us. A few years later other ships show up, destroy the first visitors and leave. Ten years later two fleets arrive.* ***Story:*** *A decade later the aliens come again, this time with a fleet of ships, and destroy the visitors and leave. One thousand years later, a new alien ship arrives, a vessel similar to the first. One hundred years later the alien ships finally come again, this time with over 500 ships, destroy the 100 ships that came the previous year, then use the surviving alien vessels to create their base*

The re-use of the term 'years later' assisted in increasing the cosine similarity F1 score. Despite the intention of maintaining a focus on rewarding semantic similarity, this shows that repetition is still also rewarded when implementing this metric. This same prompt / story pair was the 15th highest rated by humans out of 100 available.

The results in Table 1 show weak correlation between human judgement scores and automated metric scores. The highest correlation between an automated and a human metric is 0.41 between story-prompt relatedness (Q1) and BERTScore-1. Regarding story-interest as defined by humans, there was a very low correlation of 0.28 with BERTScore-1 and 0.23 with BERTScore-2. There was almost no correlation with the BERT-NSP scores; this metric found for most cases that sentences logically followed each other, however it did not provide the more granular level of analysis that human surveying and BERTScore provided.

4.4 Discussion

The machine-generated stories tend to match the theme/genre of the prompts. If this theme is not of any interest to the evaluator, they may mark this story-interest score low on the scale. Strong correlation between prompt interest and story interest however was not observed. In general, the prompts were quite specific. They set a certain tone or introduced a theme that defined a direction that a story 'should' take. Whilst this was still an open-ended style task, vaguer, less specific prompts may provide more leeway for the model to produce stories that

humans would deem relevant. While some stories generated could be considered related to the prompt, the story may have taken a secondary semantic element of the prompt to build upon, rather than use the predominant or primary theme or idea.

The last question in the survey invited the survey participants to comment on the stories. The following is a summary:

- They found the themes somewhat unsettling.
- Some stories did not make sense.
- Some stories came across as poetic, but also noted that this may have been a coincidence or fluke.
- The stories came across like 'blurbs that would be seen on the back of a book cover'.
- The stories sometimes went in a direction that human stories would not, which generated interest.
- The stories fail to continue expanding on the most interesting part of the prompt.
- There are some funny generations, there is also a surreal aspect to some of them, and even some that are profound (e.g., "I gave you all you gave me").
- The style seemed different to human writing, although this wasn't necessarily bad
- The stories written by the computer were sometimes more abstract (than the prompts)

The BERT-NSP results suggest that the sentences generated follow a logical order. BERTScore and BERT-NSP scoring is undertaken at a sentence level and aggregated for each story, whereas the human evaluators were asked to judge the story in its entirety. This is relevant, as BERTScore scores may be impacted by one or two low results in a sentence-pair within a story, thereby lowering the overall average score for an otherwise strong story. Regarding human evaluation and automated metric comparisons, the survey question that BERTScore-1 correlates with most closely – albeit with a low positive correlation of 0.41 – is story-prompt relatedness, which aligns with what BERTScore-1 is trying to achieve: semantic relatedness between the prompt and each story-sentence.

5 Conclusion

Whilst it is established that modern transformer models generate significantly more fluent text than their predecessors, evaluation of narrative elements of their output continues to be a challenge. Many standard automated evaluation metrics exist for text generation that reward repetition of the input; this is not a success metric in narrative text generation. Our survey results show a strong correlation between story coherence and story interest. Given that the average interest scores were low, this suggests perhaps that the GPT-Neo model does not always output coherent stories. There is a fine balance to suspending disbelief in storytelling, and machine-generated text is shown in this study to

often lack this level of nuance. When the text has close to human fluency, this raises the human evaluator's expectations; the machine-generated stories must be as interesting and coherent as every-day human-generated stories are to satisfy human evaluators.

Future work in the area could involve identifying story prompts for a certain specific genre (crime for example) and generate stories consistent with this genre. If evaluators with an interest this genre were recruited, this may reduce the occurrence of low scores due simply to evaluators' lack of interest in the topic, regardless of the quality of the output. There is also scope, given a sufficiently high volume of human judgements, to train a new evaluation system, allowing for the development of an automated metric fit for evaluating narrative text.

From a narrative perspective, the stories generated by GPT-Neo leave us somewhat short in terms of consistently providing interest; their success is somewhat hit-and-miss. More immediate consistent success for these systems may be achieved through generating non-narrative-style text, or by employing a hybrid machine-human approach.

References

1. Akoury, N., Wang, S., Whiting, J., Hood, S., Peng, N., Iyyer, M.: STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6470–6484. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.525, `https://aclanthology.org/2020.emnlp-main.525`
2. Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S.: GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow (2021), `http://github.com/eleutherai/gpt-neo`
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. ArXiv **abs/2005.14165** (2020)
4. Celikyilmaz, A., Clark, E., Gao, J.: Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799 (2020)
5. Chaganty, A.T., Mussman, S., Liang, P.: The price of debiasing automatic metrics in natural language evaluation. arXiv preprint arXiv:1807.02202 (2018)
6. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., Smith, N.A.: All that's' human'is not gold: Evaluating human evaluation of generated text. arXiv preprint arXiv:2107.00061 (2021)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Fan, A., Lewis, M., Dauphin, Y.: Hierarchical neural story generation. arXiv preprint arXiv:1805.04833 (2018)
9. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)

10. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research **61**, 65–170 (2018)
11. Hashimoto, T.B., Zhang, H., Liang, P.: Unifying human and statistical evaluation for natural language generation. arXiv preprint arXiv:1904.02792 (2019)
12. Howcroft, D.M., Belz, A., Clinciu, M.A., Gkatzia, D., Hasan, S.A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., Rieser, V.: Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In: Proceedings of the 13th International Conference on Natural Language Generation. pp. 169–182 (2020)
13. van der Lee, C., Gatt, A., van Miltenburg, E., Krahmer, E.J.: Human evaluation of automatically generated text: Current trends and best practice guidelines. Comput. Speech Lang. **67**, 101151 (2021)
14. van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., Krahmer, E.J.: Best practices for the human evaluation of automatically generated text. In: INLG (2019)
15. Lowe, R., Noseworthy, M., Serban, I.V., Angelard-Gontier, N., Bengio, Y., Pineau, J.: Towards an automatic turing test: Learning to evaluate dialogue responses. arXiv preprint arXiv:1708.07149 (2017)
16. McIntyre, N., Lapata, M.: Learning to tell tales: A data-driven approach to story generation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 217–225 (2009)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
18. Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., Harchaoui, Z.: Mauve: Measuring the gap between neural text and human text using divergence frontiers. Advances in Neural Information Processing Systems **34** (2021)
19. Purdy, C., Wang, X., He, L., Riedl, M.: Predicting generated story quality with quantitative measures. In: Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference (2018)
20. Roemmele, M., Gordon, A.S., Swanson, R.: Evaluating story generation systems using automated linguistic analyses. In: SIGKDD 2017 Workshop on Machine Learning for Creativity. pp. 13–17 (2017)
21. Sellam, T., Das, D., Parikh, A.P.: Bleurt: Learning robust metrics for text generation. In: ACL (2020)
22. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. ArXiv **abs/1706.03762** (2017)
23. Yao, L., Peng, N., Weischedel, R.M., Knight, K., Zhao, D., Yan, R.: Plan-and-write: Towards better automatic storytelling. ArXiv **abs/1811.05701** (2019)
24. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. ArXiv **abs/1904.09675** (2020)
25. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Rescaling bertscore with baselines. `https://github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md` (2020)