Predicting Generated Story Quality with Quantitative Measures

Christopher Purdy, Xinyu Wang, Larry He, & Mark Riedl
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA, USA
{cpurdy3,xinyuwang,larry.he,riedl}@gatech.edu

Abstract

The ability of digital storytelling agents to evaluate their output is important for ensuring high-quality human-agent interactions. However, evaluating stories remains an open problem. Past evaluative techniques are either model-specific—which measure features of the model but do not evaluate the generated stories—or require direct human feedback, which is resource-intensive. We introduce a number of story features that correlate with human judgments of stories and present algorithms that can measure these features. We find this approach results in a proxy for human-subject studies for researchers evaluating story generation systems.

Introduction

Narrative is an important tool that humans use to convey information between one another. It allows people to recount previous experiences, entertain an audience, and describe the world. If we want digital agents to be able to communicate with humans, it follows that these agents should be able to tell narratives that are relatable to humans. However, low-quality stories are not likely to engage human audiences. It is thus in the best interest of developers of story generators to estimate how effective these systems actually are.

Researchers commonly apply human-subject studies to evaluate generator quality. Researchers can take a number of generated stories and ask humans to judge their features, e.g. grammaticality or novelty. The judgments reported by these subjects can be used to estimate the quality of the generator: if the examined stories are judged as poor, there may be deficiencies with the generator.

The human-subject study is the gold standard for evaluating narrative generation systems, but such protocols are costly to run in terms of time, human resources, and participant compensation. Because of these constraints, this evaluation methodology is not conducive to rapid prototyping or model calibration. While humans are currently the best-equipped to evaluate story quality, it would be beneficial for researchers to also have reliable *proxies* for human judgments of story quality. Researchers can use these proxies to rapidly iterate on generation algorithms and save human-subject studies until necessary. In this work, we define a set of quantitative measures that collectively serve as such a proxy and validate them against human subjective ratings of story quality.

We describe two main contributions: (1) We describe four story features—grammaticality, temporal ordering, local contextuality, and narrative productivity—that correlate with human judgments of story quality. (2) We provide and evaluate quantitative measures—algorithms that can be run—that correlate with human judgments of the above story features.

Background and Related Research

Automated Story Generation has been a research problem with a long history. The most popular techniques have been symbolic and logical planning (Meehan 1977; Lebowitz 1987; Cavazza, Charles, and Mead 2002; Pérez y Pérez and Sharples 2001; Porteous and Cavazza 2009; Riedl and Young 2010; Farrell and Ware 2016), case-based and analogical reasoning (Turner 1994; Gervás et al. 2005; Ontanón and Zhu 2010). Machine learning has been used to attempt to learn story domain models or to identify segments of story content available in an existing repository to assemble stories (Swanson and Gordon 2012; Li et al. 2013). Recurrent neural networks attempt to model story progression by learning from large-scale textual corpora (Martin et al. 2018; Khalifa, Barros, and Togelius 2017; Fan, Lewis, and Dauphin 2018; Roemmele 2018; Clark, Ji, and Smith 2018). Most story generation systems are capable of producing stories of only a few sentences to a paragraph.

One of the most commonly used evaluation techniques in story generation domain is the human-subject study (Lukin, Reed, and Walker 2017; Li et al. 2013; Zhu and Ontanón 2013). Human-subject studies typically involve the subjects interacting with the story generation system or with the generated stories of the system. However, there are numerous other existing evaluation techniques that do not require direct human involvement.

Previous researchers have used quantitative metrics to predict the quality of machine learning-based generative systems, including those for neural machine translation (Bahdanau, Cho, and Bengio 2014; Cho et al. 2014), story generation (Lukin, Reed, and Walker 2017; Martin et al. 2018; Fan, Lewis, and Dauphin 2018), and image captioning (Vinyals et al. 2015; Xu et al. 2015). For example, *per-*

plexity is a measure of the ability of a probabilistic system to predict proper output as compared to a gold standard test dataset (Jelinek et al. 1977). It is often used as a measure of the quality of generated text, but one major limitation is that this score does not evaluate the quality of any given output produced by a generative model: it describes the model as a whole and its ability to reproduce a given corpus. This is problematic because it does not consider the possibility that novel stories can be valuable.

Another common machine learning metric is the BLEU score (Papineni et al. 2002) that evaluates generated text based on the word overlap with a ground truth sequence. Like perplexity, it assumes an *a priori* known correct answer, which does not factor in the possibility of correct but novel generation. This evaluative method also is not grounded in human judgment: a high-scoring story can still be of poor quality depending on the ground-truth used.

Numerous methods of writing evaluation exist in linguistics that rely on formulas using statistical properties of the words and sentences of passages. These methods include reading ease and lexical complexity. Reading ease describes how difficult it is to read a passage and is commonly expressed as the minimum education level needed to read the passage (Flesch 1948) (Kincaid et al. 1975). Lexical complexity measures the breadth of language in a passage (Richards 1987). Automated essay scoring systems incorporate lexical complexity and reading ease with additional metrics to grade student writing. These metrics include grammaticality and essay structure (Attali and Burstein 2004). Machine learning models learn to predict scores from the selected measurements (Alikaniotis, Yannakoudakis, and Rei 2016). While these metrics are useful in specific domains, little research has been done to see if these techniques can be adapted to predict story quality.

Selecting and Quantifying Story Features

The previously described methods either do not directly evaluate generated stories or did so in ways that are not grounded in human judgment. To overcome this shortcoming, we built an evaluative methodology that more closely mirrors human practice. To bridge the gap between machine evaluation and human intuition, we identified specific features that we hypothesized would correlate with actual human judgments of story quality: grammaticality, temporal ordering, narrative productivity (reading ease and lexical complexity combined), and local contextuality. We describe these features and their quantitative instantiations below.

Grammaticality

Grammaticality indicates how well language in the story adheres to rules of grammar. We evaluate grammaticality using a grammar rating system, borrowing methodology from Heilman et al. (2014). Specifically, we use ridge-regression train a grammar model from spelling and n-gram features using the "Grammatical versus Ungrammatical" (GUG) dataset. The features we isolate to train this grammar model are:

- The number and proportion of misspelled words in the sentence
- The max-log and min-log probabilities of the count of n-grams in the sentence, according to English Gigaword (fifth edition), for $n \in [1, 3]$

The GUG dataset provides English-language learner sentences and human grammaticality judgment labels. For example, a sentence like "He is only a little boy do not everything clearly?" is annotated with the lowest ordinal value of 1 ("Incomprehensible"), while the sentence "I stayed in a dorm when I went to college" is annotated with the highest ordinal value of 4 ("Perfect"). The model trained on this dataset using the above features is used to predict human ordinal judgments of grammaticality on a [1, 4] ordinal scale. This prediction is what we label as the "grammaticality" story feature.

Narrative Productivity

We define narrative productivity as a general measure of language complexity. Narrative productivity is measured with a variety of metrics directly taken from linguistics research in evaluating writing, which we group under two umbrellas: reading ease and lexical complexity. We measure lexical complexity with (a) *type token ratio* and (b) *corrected type token ratio* (Richards 1987) (Hess, Sefton, and Landry 1986). We measure reading ease with Flesch Reading Ease (Flesch 1948), Flesch-Kincaid Grade Level (Kincaid et al. 1975), and SMOG Index (Mc Laughlin 1969). We collect a number of features separately here in order to capture multiple dimensions of language complexity simultaneously.

Local Contextuality

Local contextuality measures the semantic relevance of sentences in the context of their neighboring sentences. Stories build upon previous ideas by introducing new concepts that sensibly further the plot. Sometimes, though, random events and concepts can confuse readers. Consider a brief romance story about two high-school friends reunited after a long separation. We can imagine two different endings for this story: (1) "Slowly, they start to fall in love," or (2) "Slowly, they develop a gambling addiction."

Both of these are valid English sentences and can occur in certain stories. In this story, however, the gambling addiction ending is a non sequitur, and humans can recognize this. In contrast, computer generated stories often include disparate terms that lack meaningful relationships between them, like those in the latter version. A digital story generator that wishes to tell convincing stories should be able to tell when the stories it is producing maintain local contextuality.

To measure local contextuality procedurally, we use one-dimensional sentence embeddings obtained from a Sent2Vec (Pagliardini, Gupta, and Jaggi 2018) model trained on the CMU Plot Summary corpus (Bamman, O'Connor, and Smith 2014). For each sentence, the Sent2Vec embedding process determines the word embeddings of each constituent unigram as well as source embeddings of its n-grams and obtains the sentence embedding via averaging. We procedurally compare the semantic contents of two input sentences

by computing the *cosine similarity* between their Sent2Vec embeddings, x and y. A cosine similarity score of 1.0 indicates complete similarity, while a score of 0.0 indicates no similarity. To compute the local contextuality score of a story, we take the average of the cosine similarities of every adjacent pair of sentences in the story.

Temporal Ordering

While local context is about whether two adjacent sentences preserve context, temporal ordering looks at the plausibility that one sentence should follow another. These sentences can be adjacent or separated by some number of other sentences. For example, consider a story that contains the following sentences: (1) "Jane ordered food," (2) "Jane texted her friend," and (3) "Jane ate her meal." It is plausible that sentence 1 precedes sentence 3 even though not adjacent, but the opposite ordering is less plausible. Events in stories have causal relations (Trabasso and van den Broek 1985; Graesser, Singer, and Trabasso 1994). However, how to automatically identify causal relations in arbitrary texts is an open question. An alternative we employ is to learn highly probable temporal patterns of behavior of story characters (e.g., the probable order of events in a restaurant scenario) and measure the occurrence of these patterns with the assumption that patterns of behavior correlate with causal relations.

We take a number of steps to identify temporal ordering in stories. Following the work of Martin et al. (Martin et al. 2018), who found that the accuracy of neural story generation systems can be improved by abstracting sentences to tuples, we convert natural-language sentences into a tuple representation (called an *event*), a tuple $\langle s, v, o, m\rangle$ where v is the root verb of the sentence, s is the verb's subject, o is the verb's direct object, and m is a modifier term (typically a preposition). For example, the sentence "John quickly locked the gold in the bank vault" would be $\langle johh, lock, gold, vault\rangle$. First, a sequence of events is extracted from a story with the Stanford CoreNLP toolkit (Manning et al. 2014). To account for the fact that a pattern of meaningful events may be dispersed across a story (i.e., not only occurring in adjacent sentences), we use a *skipping recurrent neural network* over the sequence of generated events to select the most important sentences.

Skipping recurrent neural networks (SRNNs) (Sigurdsson, Chen, and Gupta 2016), which were first applied to photo album summarization, select a subset of k elements from a sequence that preserve the information of the whole sequence as much as possible. The network is referred to as "skipping" because the selected entities do not need to be adjacent to each other in the input sequence. We use an adaptation of the SRNN architecture that operates on stories for the task of story summarization (Harrison, Purdy, and Riedl 2017). Instead of training the SRNN on pixels, images, and albums we train on words, sentences, and stories from the CMU Plot Summary corpus (Bamman, O'Connor, and Smith 2014).

SRNNs, in the context of temporal ordering measures for stories, extract the pivotal events of a story based on patterns of events shared across many stories. These events reflect the general structure of the story it summarizes. In the case of romance stories, a recurring structure is "boy and girl meet" and then "boy and girl fall in love." The SRNN, after having seen many stories with this pervasive formula, learns that these events are integral to the story together in this order. Other embellishing details, such as "boy and girl live in San Diego", are discarded.

For each story in the CMU Plot Summary Corpus, we use the SRNN to select the two most important sentences from that story and then select the root verbs, v_1 and v_2. If v_1 occurs before v_2, we infer the temporal relation, $v_1 \rightarrow v_2$. We use the set of temporal relations to construct a directed graph. These inferred relations are transitive; if we observe $v_1 \rightarrow v_2$ and $v_2 \rightarrow v_3$, we infer that $v_1 \rightarrow v_3$ is also a legal relation. Furthermore, if we observe both $v_1 \rightarrow v_2$ and $v_2 \rightarrow v_3$, then both relations are legal.

We refer to this as a temporal order model. We show a simplified model for visualization purposes in Figure 1.

To measure the temporal ordering of an input story, we examine each pair of sentences in that story, extract the root verbs, and check to see if they are connected in the temporal relation model graph. If both verbs are present in the graph, connected by a temporal relation, and occur in the same order that the directed relation designates, then the sentences are deemed as properly ordered. The temporal ordering score for the story is the number of properly ordered pairs divided by the total number of pairs for which both verbs appear in the graph.

Evaluation

To create a set of measures that can act as proxies for human-subject studies, we pose two hypotheses:

H1. The selected story features correlate with human judgments of story quality.

H2. Our measures correlate with the selected story features.

To test these hypotheses, we prepared a corpus of short stories, controlling for grammar, local contextuality, temporal ordering and narrative productivity. We conducted a human-subject study in which participants read short, six-sentence stories and provided qualitative feedback about the features as well as overall quality and enjoyment. Finally, we ran our automated measures on the same corpus of stories and compared the scores from our algorithms against human-reported scores.

Corpus Preparation

We randomly sampled three stories from the CMU Plot Summary corpus (Bamman, O'Connor, and Smith 2014). We ensured that each of these stories had six sentences to control for story length as well as matched the target length of many existing story generation techniques. We selected six-sentence stories to reflect the current state of story generation research focusing on short text segments and also to account for cognitive load during the human-subject study.

For each of these base stories, we applied changes of different magnitude to create isolated deviations to our target features. This resulted in nine categories of stories:

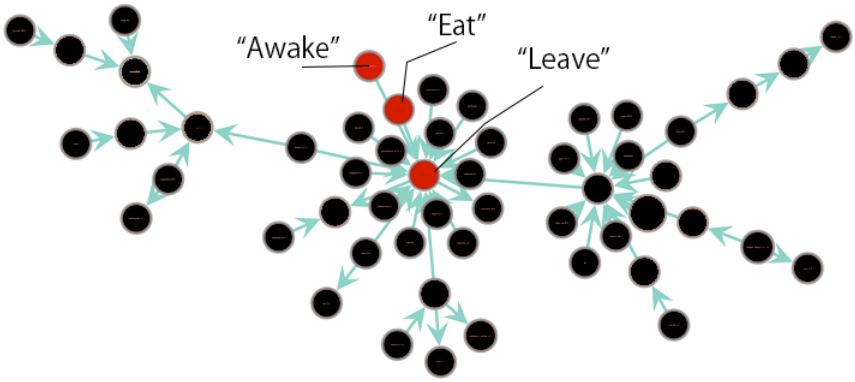

Figure 1: A simple temporal ordering network. Each node represents an event and each directed edge represents a temporal ordering.

1. Three base stories, each with six sentences.

Grammatical interventions:

2. The three stories from (1), each with with a small number of grammatical errors (typos and deleted grammatical tokens).
3. The three stories from (1), each with a high number of grammatical errors (typos and deleted grammatical tokens).

Temporal Ordering interventions:

4. The three stories from (1), each with two events randomly shuffled.
5. The three stories from (1), each with multiple events randomly shuffled.

Local Contextuality interventions:

6. The three stories from (1), each with a small number of noun substitutions (e.g. replacing "store" with "volcano")
7. The three stories from (1), each with a large number of noun substitutions (e.g. replacing "store" with "volcano")

Narrative Productivity interventions:

8. The three stories from (1), each with a small number of clause simplifications and redundant clauses added
9. The three stories from (1), each with a large number of clause simplifications and redundant clauses added.

We introduced different magnitudes of interventions in order to ensure that our predictive measures could apply to stories of different qualities. For example, if we only looked at stories with bad grammar, we would have no idea how well our measures could predict the quality of stories with only a handful of mistakes. When we apply multiple degrees of intervention for each intervention type, we increase the generalizability of our findings. We examined a total of 27 stories, given we make use of three for each of the nine groups.

Methodology

We recruited 500 participants from Amazon's Mechanical Turk crowdsourcing platform. Each participant was asked to read a single story from our corpus. Each participant was asked to state their level of agreement with the following statements on a scale from 1 to 5 (with a score of "5" indicating complete agreement and a score of "1" indicating complete disagreement):

- This story exhibits CORRECT GRAMMAR.
- This story's events occur in a PLAUSIBLE ORDER.
- This story's sentences MAKE SENSE given sentences before and after them.
- This story AVOIDS REPETITION.
- This story uses INTERESTING LANGUAGE.
- This story is ENJOYABLE.
- This story is of HIGH QUALITY.

The first three questions correspond to grammaticality, temporal ordering, and local contextuality, respectively. Since the term "narrative productivity" is not easy to ask about in a single question, we have two questions that deal with individual facets of narrative productivity: repetition and interesting language. Participants were paid $5.00 for completion of the task, which took on average 15 minutes.

Table 1: The correlation coefficients from each of the story features compared against story *quality*.

Story Feature	ρ **correlation**
Grammaticality	0.405
Temporal Ordering	0.431
Local Contextuality	0.552
Repetition	0.112
Interesting Language	0.532

Results

We ran two separate analyses, each answering one of our hypotheses. All values reported are statistically significant at $p < 0.05$.

Table 2: The correlation coefficients from each of the story features compared against story *enjoyment*.

Story Feature	ρ **correlation**
Grammaticality	0.249
Temporal Ordering	0.446
Local Contextuality	0.490
Repetition	0.105
Interesting Language	0.430

Table 3: The correlation coefficients from natural language metrics compared against *language interestingness*.

Measure	ρ **correlation**
Type-token ratio	0.237
Corrected type-token ratio	0.271
Flesch Readability	-0.325
Flesch-Kincaid	0.334
SMOG Index	0.305

Relationship between Features and Story Quality

To test our first hypothesis—that the story features correlate with perceptions of story quality—we compared the responses to each of the first five questions of the survey with the responses to the story quality question. We used Spearman's rank-order correlation test to determine the strength and correlation between perceived story quality and all of our proposed story features. The correlation coefficient for this test, ρ, is bounded in the range $[-1, 1]$, with a value of 1 corresponding to perfect positive correlation, a value of -1 corresponding to perfect negative correlation, and a value of 0 corresponding to no correlation. The correlation coefficients for each of these tests can be seen in Table 1. We also ran the same procedure comparing the selected story features against story enjoyability, the results of which are reflected in Table 2.

Local contextuality and interesting language are found to be strongly correlated (in Spearkman's rank-order tests, a ρ of 0.5 is generally considered strong correlation) with perceived story quality and enjoyability. Grammaticality and plausible temporal ordering slightly less strongly correlated with story quality. Repetition appears to be weakly correlated with story quality, but this may be due to the overall short lengths of the presented stories; repetition in a story of only six sentences may be less detrimental than repetition in a story of 600 sentences. Overall, we find the results provide evidence that measuring these features is worthwhile for the sake of generating stories.

Automated Measures

To test our second hypothesis—that our automated measures correlate with human judgments—we compare the scores our algorithms produce to the scores from the human-subject study. Specifically, we look at rank-order of human scores and algorithmically generated scores. We expect that when stories are ranked using numerical scores produced by humans, the same ranking appears when using algorithmically produced scores.

Grammaticality Using story categories 1, 2, and 3 as a testing set, we ran the Spearman's rank-order correlation test between the predicted scores from our grammar model and "grammaticality" results from the survey. There was moderate correlation between the two groups, with a ρ correlation of 0.386. This shows that, with a grammar model trained on more features, we can build an evaluator with judgments closer to human intuition of grammaticality of stories.

Narrative Productivity Using story categories 1, 8, and 9 as a testing set, we ran the Spearman's rank-order correlation test between numerous lexical complexity and reading ease measures and the "interesting language" and "avoiding repetition" questions from the survey. These results are shown in Tables 3 and 4.

The lexical complexity and readability formulae we implemented have moderate correlation to the survey results for the "avoiding repetition" and "interesting language" questions. Researchers should consider each of these when examining grammaticality. It is likely that more complex language measures will be required in order to capture the messiness of perceived narrative productivity.

Local Contextuality Using story categories 1, 6, and 7 as a testing set, we ran the Spearman's rank-order correlation test between the Sent2Vec cosine similarity scores and "local contextuality" results from the survey. The ρ correlation was 0.490. This indicates that the Sent2Vec measure is a good way to capture differences in context between adjacent sentences. Specifically, as the average cosine similarity of the Sent2Vec measure increases, the reader's perception of the local contextuality also increases.

Temporal Ordering Using story categories 1, 4, and 5 as a testing set, we ran the Spearman's rank-order correlation test between the temporal ordering Bayesian estimate and "temporal ordering" results from the survey. The ρ correlation was 0.103. This is interpreted as a weak correlation. While there is some indication that our temporal ordering algorithm is able to predict when humans also agree with temporal orderings to some degree, this automated measure is much more noisy.

There are a number of potential causes for this. It could be symptomatic of the nature of story ordering; swapping random events in a story does not necessarily invalidate the story. The error rate in our SRNN may be too high. This is difficult to assess since there is no ground-truth for the summarizations it produces. Our simplifying assumption of only looking at verb orderings may introduce error as well. While a low correlation is still a positive result, it indicates a need for further research into algorithms capable of predicting the quality of temporal ordering.

Discussion

We have identified and justified specific story features that correlate with human judgments of story quality. The list of features is not exhaustive; the methodology presented in this

Table 4: The correlation coefficients from natural language metrics compared against *absence of narrative repetition*.

Metric	ρ **correlation**
Type-token ratio	0.474
Corrected type-token ratio	0.307
Flesch Readability	-0.326
Flesch-Kincaid	0.376
SMOG Index	0.362

paper also provides a template for us and others to add to the list of proxy measures.

Furthermore, we have quantified each of these story measures so that researchers can isolate specific strengths or weaknesses of their story generators without having to rely exclusively on human-subject studies. No automated measures cannot perfectly predict human judgments of story quality or enjoyment, but these automated proxies can provide a useful tool during the development of story generation systems. A question that often arises during development is whether a modification to a story generation algorithm has improved the performance of the system or not. The appropriate use of the automated measures is to compare two versions of the same story generation approach.

Story generation systems that rely on heuristics or scoring functions of their own output can also incorporate these metrics into their generation loop. This can be especially useful for story generation systems that treat the story generation problem as a planning and/or optimization problem.

By reporting on four different features, the automated metrics give a more fine-grained analysis of a story generation system's outputs, helping researchers understand the trade-offs between the different metrics. One may also choose to sum them together or to look at only the features that are most relevant to their system.

Identifying whether a story has a plausible temporal ordering or whether the events in a story support common causal relations is a hard, open research problem. Recent experiments (anonymized, under review) show neural networks are only able to reconstruct the correct ordering of randomized story sentences $\sim 20\%$ of the time. Thus is it not surprising that the temporal ordering metric had the lowest correlation with human rankings. However, since story generation systems are often capable of generating many thousands of stories (especially in the case of machine learning generators), one may construct experiments evaluating large numbers of outputs of two story generation systems where the law of large numbers overcomes the fact that temporal ordering scores rankings are only weakly correlated with human rankings.

Conclusions

The gold standard for the evaluation of story generation systems will always be some form of human subject study. However, human subject studies are costly and thus cannot be conducted frequently. This presents a bottleneck to AI research on story generation where one may want to make incremental adjustments to an algorithm but cannot know whether those adjustments improve the system. Most automated evaluation methods may indicate certain properties of a model or algorithm without any known correlation to human judgments.

We have shown that grammaticality, local contextuality, temporal ordering, and narrative productivity features correspond to human judgments of narrative quality and enjoyment. We have further developed algorithms for scoring each of these features and shown that the correlate with human judgments of these features. These results in conjunction suggest that we can use our metrics as a proxy for expensive and time-consuming human-subject studies of narrative generation systems. One may further be able to incorporate the metrics directly into the generation loop of story generation systems. The implications are that we can more rapidly iterate on story generation systems, scaling and speeding up research.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1350339. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Alikaniotis, D.; Yannakoudakis, H.; and Rei, M. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Attali, Y., and Burstein, J. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series* 2004(2).

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bamman, D.; O'Connor, B.; and Smith, N. A. 2014. Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 352.

Cavazza, M.; Charles, F.; and Mead, S. 2002. Planning characters' behaviour in interactive storytelling. *Journal of Visualization and Computer Animation* 13:121–131.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Clark, E.; Ji, Y.; and Smith, N. A. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. *ArXiv e-prints*.

Farrell, R., and Ware, S. 2016. Fast and diverse narrative planning through novelty pruning. In *Proceedings of the 12th AAAI International Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology* 32(3):221.

Gervás, P.; Díaz-Agudo, B.; Peinado, F.; and Hervás, R. 2005. Story plot generation based on CBR. *Journal of Knowledge-Based Systems* 18(4–5):235–242.

Graesser, A.; Singer, M.; and Trabasso, T. 1994. Constructing inferences during narrative text comprehension. *Psychological Review* 101(3):371–395.

Harrison, B.; Purdy, C.; and Riedl, M. O. 2017. Toward automated story generation with markov chain monte carlo methods and deep neural networks. In *Proceedings of the 2017 Workshop on Intelligent Narrative Technologies*.

Heilman, M.; Cahill, A.; Madnani, N.; Lopez, M.; Mulholland, M.; and Tetreault, J. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, 174–180.

Hess, C. W.; Sefton, K. M.; and Landry, R. G. 1986. Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research* 29(1):129–134.

Jelinek, F.; Mercer, R. L.; Bahl, L. R.; and Baker, J. K. 1977. Perplexitya measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62(S1):S63–S63.

Khalifa, A.; Barros, G. A. B.; and Togelius, J. 2017. DeepTingle. *ArXiv e-prints*.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command.

Lebowitz, M. 1987. Planning stories. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society*.

Li, B.; Lee-Urban, S.; Johnston, G.; and Riedl, M. 2013. Story generation with crowdsourced plot graphs. In *AAAI*.

Lukin, S. M.; Reed, L. I.; and Walker, M. A. 2017. Generating sentence planning variations for story telling. *arXiv preprint arXiv:1708.08580*.

Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Martin, L. J.; Ammanabrolu, P.; Hancock, W.; Singh, S.; Harrison, B.; and Riedl, M. O. 2018. Event representations for automated story generation with deep neural nets.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading* 12(8):639–646.

Meehan, J. R. 1977. TALE-SPIN: An interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 91–98.

Ontanón, S., and Zhu, J. 2010. Story and text generation through computational analogy in the Riu system. In *Proceedings of 6th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Pérez y Pérez, R., and Sharples, M. 2001. MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence* 13:119–139.

Porteous, J., and Cavazza, M. 2009. Controlling narrative generation with planning trajectories: the role of constraints. In *Proceedings of the 2nd International Conference on Interactive Digital Storytelling*, 234–245.

Richards, B. 1987. Type/token ratios: What do they really tell us? *Journal of child language* 14(2):201–209.

Riedl, M. o., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39:217–268.

Roemmele, M. 2018. *Neural Networks for Narrative Continuation*. Ph.D. Dissertation, University of Southern California.

Sigurdsson, G. A.; Chen, X.; and Gupta, A. 2016. Learning visual storylines with skipping recurrent neural networks. In *European Conference on Computer Vision*, 71–88. Springer.

Swanson, R., and Gordon, A. 2012. Say Anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Transactions on Interactive Intelligent Systems* 2(3):16:1–16:35.

Trabasso, T., and van den Broek, P. 1985. Causal thinking and the representation of narrative events. *Journal of Memory and Language* 24:612–630.

Turner, S. R. 1994. *The Creative Process: A Computer Model of Storytelling*. Lawrence Erlbaum Associates.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 3156–3164. IEEE.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.

Zhu, J., and Ontanón, S. 2013. Evaluating analogy-based story generation: An empirical study. In *Proceedings of the Ninth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.