Proceedings of the Fifteenth AAAI Conference on Artificial
Intelligence and Interactive Digital Entertainment (AIIDE-19)

Story Quality as a Matter of Perception: Using Word Embeddings to Estimate Cognitive Interest

Morteza Behrooz
morteza@ucsc.edu
University of California Santa Cruz
1156 High St. Santa Cruz, California 95064

Justus Robertson
jjrobert@ncsu.edu
North Carolina State University
Raleigh, NC 27695

Arnav Jhala
ahjhala@ncsu.edu
North Carolina State University
Raleigh, NC 27695

Abstract

Storytelling is a capable tool for interactive agents and better stories can enable better interactions. Many existing automated evaluation techniques are either focused on textual features that are not necessarily reflective of perceived interestingness (e.g. coherence), or are domain-specific, relying on a priori semantics models (e.g. in a game). However, the effectiveness of storytelling depends both on its versatility to adapt to new domains and the perceived interestingness of its generated stories. In this paper, drawing from cognitive science literature, we propose and evaluate a method for estimating cognitive interest in stories based on the level of predictive inference they cause during perception.

Introduction and Background

Storytelling is arguably the most powerful form of human communication. Stories have deep roots in our cultures, shape the way we develop most forms of entertainment, and play a crucial role in our social interactions. Many societal and historic reasons exist that can explain the prevalence of storytelling in human life. The most dominant reason, however, has roots in human cognition.

As we develop intelligent agents that are capable of interacting with humans in increasingly natural and social ways, it is crucial to improve their storytelling abilities. For many types of interactive agents (e.g. social robots, game characters or voice assistants) and in various contexts of interaction (e.g. entertainment, service, health care, or education) storytelling can enhance or aid the interaction by increasing engagement (Battaglino and Bickmore 2015), rapport (Bickmore and Cassell 1999), closeness (Coon, Rich, and Sidner 2013), character believability (Riedl and Young 2005; Gomes et al. 2013), perceived sociability, or ludic values, among others.

Better stories can create better interactions. However, certain intuitive and commonplace human behaviors are often deceptively difficult to recreate, as years of cultural and cognitive evolution has perfected our intuition-based way of evaluating them. For instance, it can be extremely difficult to recreate a simple nodding behavior that could seem perfectly natural to humans (Admoni and Scassellati 2017). Similarly, generating stories that are perceived to be as interesting as the stories told by humans is a difficult task to perfect, especially when this task is meant to serve various contexts of interaction and domains of storytelling. Towards the goal of creating better stories, and as story generation techniques improve and their use-cases expand, the question of evaluating the generated stories gains more significance.

Researchers have often relied upon human-subjects to evaluate a story generator. While this approach remains a gold standard, having access to automated evaluation techniques is beneficial for two reasons. Firstly, as human-subject studies can be costly and time-consuming, an automated evaluation can be performed much more frequently (e.g. for prototyping or fine-tuning machine learning models). Secondly, operating in different contexts and domains may change the evaluation criteria, and an automated measure, if informed of such changes, can be capable of adjusting itself. For instance, in (Sidner 2015), researchers introduce an agent that changes its behavior based on how much rapport it has built with a long-term companion.

A story generator's intended use-case and narrative delivery paradigm have important implications for the constraints of the generation techniques, as well as the automated evaluation metrics it can have. Experiences such as games which involve a mesh of different forms of entertainment have been a great motivation for story generation. Thus, many story generators have been developed in the context of one game, e.g. in (McCoy et al. 2011), or otherwise one fixed domain, e.g. in (Elson 2012). As such, judging the quality of their generated stories can fully or partially depend on the semantics of that particular domain. This a priori knowledge can help in evaluating the quality of the generated stories, both at the fabula level (e.g. by the way of knowing the significance of events) and the narrative level (e.g. by a more informed choice of words or event ordering).

Interactive agents, however, are increasingly operating across multiple domains and contexts. Hence, in such cases, relying on a priori domain semantics is not a viable option, neither for the generation nor the automated evaluation processes. Generating stories without reliance on domain knowledge is sometimes called "open story generation" (Martin et al. 2017). Generic evaluation metrics that are often used in such systems are discussed in the next section.

Story Evaluation Metrics

Story generation, as a field of research, has had a longer history than many of the methods it has used. The most popular techniques include various forms of planning (Turner 1994; Meehan 1977; Lebowitz 1987; PÉrez and Sharples 2001; Porteous, Cavazza, and Charles 2010; Riedl and Young 2010; Farrell and Ware 2016) and case-based and analogical reasoning (Gervás et al. 2005; Ontañón and Zhu 2010; Turner 2014). Other methods have relied on crowd-sourcing or textual corpora to model story domains (Swanson and Gordon 2008; Li et al. 2013). Recurrent neural networks (RNNs) and language models have also been used more recently to generate stories using models trained on large corpora of stories or text (Martin et al. 2017; Khalifa, Barros, and Togelius 2017; Fan, Lewis, and Dauphin 2018; Clark, Ji, and Smith 2018). Most recently, reinforcement learning has been used to control the generation of stories using neural networks by assigning a goal to this process (Tambwekar et al. 2018a). All of these approaches, especially those not limited to a particular domain (e.g. RNNs and language models), would benefit from automated evaluation metrics.

Automatic assessment of the quality of the generated text is one of the main evaluation metrics in natural language processing and generation. Scores such as BLEU (Papineni et al. 2002) and PINC (Chen and Dolan 2011) evaluate the quality of the generated text against a ground-truth source (in tasks such as translation or paraphrasing). Other scores target more generic concepts of coherence and cohesion in text (Foltz, Kintsch, and Landauer 1998; Graesser et al. 2004). Perplexity is another metric for evaluating the model with which text is being generated (Jelinek et al. 1977), although it does not evaluated the generated text directly.

When not using the gold standard of human-subjects, story generation research has used such metrics to evaluate the generated stories (Fan, Lewis, and Dauphin 2018; Lukin, Reed, and Walker 2017; Martin et al. 2017). However, while these scores can provide some estimation of the quality of the generated text, they are 1) not suited for many machine learning approaches to story generation (Purdy et al. 2018), and 2) they do not focus on what makes stories compelling and interesting; an important consideration for stories used in interactions.

In a recent and most relevant work, Purdy et al. introduce four quantitative story quality metrics to address these issues (Purdy et al. 2018). These measures can be used to evaluate a generated story, and are shown to correlate with human judgements of narrative quality and enjoyment, hence acting as "proxy measures". These measures are:

- Correct spelling and grammar use (**"grammaticality"**),
- Linguistics-based measures of reading ease and language complexity (**"narrative productivity"**),
- Semantic similarity of adjacent sentences (**"local contextuality"**), and,
- Level of adherence to the usual ordering of events in stories, e.g. "eat" after "order" (**"temporal ordering"**).

While the first two measures are strictly focused on the use of language, the last two focus on more semantic evaluations. Local contextuality, as defined above, uses sentence embeddings (Pagliardini, Gupta, and Jaggi 2017) to estimate semantic coherence and investigate whether sentences are relevant to each other in their progression. Temporal ordering investigates if the verbs in a story adhere to an ordering network of precedence rules built from many stories seen before, a network similar to Plot Graphs (Li et al. 2013).

While Purdy et al. find correlations between their proxy measures and "enjoyment" in human subjects, it is arguable that the main source of enjoyment in the perception of a story comes from finding it as *interesting*. Indeed, if a story has spelling errors, is hard to read, contains irrelevant sequence of sentences or unreasonable verbs, perceiving it would be a much less enjoyable experience than perceiving one free of those problems. However, a story that observes all such measures may still be boring and mundane. Hence, while these proxy measures have inspired us in our direction of research, we seek to expand them to other areas. One such area is the perceived story interestingness, and particularly, cognitive interest.

Story Interestingness

There are many reasons why a story might be perceived as interesting to an audience, many of which may be categorized as subjective or may have roots in culture and environment. Moreover, the art of authorship and telling of a story is often a source of interestingness and such art can be contextual, nuanced and subtle. Nonetheless, there exists a history of a long effort, both in the story generation community and in the cognitive science and cognitive psychology, to understand what causes interest in stories.

Categorizations of Interest. Most researchers have categorized one's possible interest in a story in two main camps based on where they estimated the source of the interest to be: a predisposition of the listener or a property of the story. These two camps have received different names by various researchers, such as *emotional* and *cognitive* (Kintsch 1980), *individual* and *situational* (Hidi and Baird 1986), *interestedness* and *interestingness* (Frick 1992) or *topic* and *cognitive* (Campion, Martins, and Wilhelm 2009). We will refer to these two camps as "emotional" and "cognitive" for simplicity. In addition, a group of *absolute interests* (e.g. danger, power, sex) are introduced by Schank (Schank 1979), and corroborated by other researchers under various names such as *generically important topics* (Freebody and Anderson 1986) and *human dramatic situations* (Wilensky 1983). A recent taxonomy of emotional interests which includes Schank's absolute interests is offered in our previous work (Behrooz et al. 2018).

Cognitive Interest. Many researchers have developed and empirically evaluated theories of the mechanisms that lead to the establishment of cognitive interest. Notable theories include *unexpectedness* (Schank 1979), interaction between background knowledge, uncertainty and *postdictability* (Kintsch 1980; Iran-Nejad 1987), *incongruity* (Mandler 1982), *change in belief* (Frick 1992), generation of *inference* (Kim 1999) and the generation of *predictive inference*

(Campion, Martins, and Wilhelm 2009). Crucially, the validity of these various theories of cognitive interest is not mutually exclusive; but rather, they often attempt to explain each other.

In this paper, we discuss a method for creating a proxy measure for cognitive interest. We will then report on a two-phased user study to evaluate this measure and will discuss the results. Our goal of contribution is to establish a connection between quantifiable story evaluation metrics and the cognitive qualities in the perception of stories, as an additional layer of story evaluation besides language use and semantic coherence.

Cognitive Interest as a Proxy Measure

As previously discussed, expanding the evaluation metrics to include a measure of interestingness is distinctly different than evaluating many of the surface features. In fact, at times a (seemingly) bad quality on the surface can contribute to cognitive interest. For instance, a story that adheres less than perfectly to the known sequence of verbs (thus obtains a mediocre value in terms of temporal ordering in (Purdy et al. 2018)), may contain an *unexpected* event that contributes to cognitive interest. However, too much deviation from known sequences of verbs is probably not very interesting either. This emergent balance is reminiscent of Kintsch's idea of cognitive interest as an inverted-U-shaped function of knowledge and uncertainty (Kintsch 1980). To this end, we believe that the addition of proxy measures of cognitive interest would be useful to the automated evaluation of generated stories.

Quantitative Estimation of Predictive Inference

We focus on predictive inference which is known to be the main cause of cognitive interest according to one of the more recent theories, and one compatible with theories before it (Campion, Martins, and Wilhelm 2009). To this end, we sought to develop a proxy measure that can estimate the generation of predictive inference in the audience's mind. The best known authorship skill that can generate predictive inference is *foreshadowing*. Indeed, not all cases of foreshadowing lead to predictive inference, as the hint provided in the story can be too subtle to drive predictive inference and a hint's connection to future events may be only revealed later in the process of postdictability. However, many cases of foreshadowing stand out to the audience as a curious case; a state of mind that is often the author's intent.

In (Bae and Young 2008), focusing on surprise as a driving factor, a planning-based framework for generating flashback and foreshadowing is provided. While this research was an inspiration for our work, we sought an approach that can estimate the presence of foreshadowing without relying on explicit a priori knowledge.

Word Embeddings

One of the products of the advancements of deep learning is the dramatic increase in the quality of word embeddings: high-dimensional vectorized representations of words largely based on co-occurrence in large corpora. This increase in quality has even opened doors to performing analogical reasoning using word vectors (Gladkova, Drozd, and Matsuoka 2016). Word2Vec (Goldberg and Levy 2014), GloVe (Pennington, Socher, and Manning 2014) and FastText (Joulin et al. 2016) are three successful models for creating word embeddings; however, they are context-independent and associate a certain word in a corpus with a single vector regardless of the sentences and contexts it appears in. Newer models, such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2018) take the context into consideration and associate the same word with different vectors based on the context it appears in (e.g., adjacent words, the containing sentences or story event).

Estimating the presence of foreshadowing without a semantic model of the domain is a complicated task. Foreshadowing can take many different shapes, be causal or non-causal, or depend on domain-specific clues. However, many cases of foreshadowing involve usage of words that co-occur in many other contexts. Thus, such words are likely to have similar word vectors in an embedding space, especially one that considers the context. This is the main intuition behind our approach.

Method

Given a short story, we first remove all of the stop words and named-entities in it. Then, using BERT (Devlin et al. 2018) embeddings pre-trained on a books corpus (Zhu et al. 2015) (with 1024 dimensions), we extract word vectors for every remaining word in the story. As previously mentioned, this model yields different vectors for each occurrence of a word in the story.

In order to simulate the linear nature of the perception of the narrative, we incorporate a concept we call "moving cosine similarity". Starting from the second sentence of the story (word location b), we calculate the cosine similarity of every word vector with the average of all of the word vectors that precede it in the story. In other words, we calculate:

$$sim(w_i) = cosine(mean([w_b..w_{i-1}]), w_i)$$

for every word w_i. Consider the example short story seen in Table 1 which contains a case of foreshadowing. Calculating the moving cosine similarities for all of the words starting from the second sentence will yield a sequence of values. A chart of these values for our example story is seen in Fig. 1. Assuming that every foreshadowing will consist of two main parts in two sentences (e.g., a "hint" and a "twist"), we notice in our example story in Table 1 that the words `distracted` and `tired` are key words of the *hint* sentence (where the waiter is distracted), and the word `wrong` is the key word of the *twist* sentence (where the food is wrong). These words show an anomalously low amount of moving cosine similarity.

In order to algorithmically find the anomalously low values in the sequence of moving cosine similarities, we use a simple outlier detection algorithm. Such algorithms usually have a threshold with which they detect outlier values. We seek to find out if there are exactly two different sentences in the story where such anomaly occurs (such as in Fig. 1)

Table 1: An example story that contains a case of foreshadowing. The words in bold correspond to the dips in the chart seen in Fig. 1.

Sam and Judy went out for dinner at their favorite restaurant. While driving to the restaurant, Judy's favorite song played on the radio. Sam found a parking space at the very front of the restaurant. Sam and Judy were seated immediately and ordered their favorite food to the waiter. He looked **distracted** and **tired** but was polite while taking their order. Sam's favorite song played on the radio while they waited for their food. When the waiter returned with their food it was all **wrong**! The waiter apologized and returned a few minutes later with the correct order. Sam and Judy enjoyed their meal. They paid their tab, left a tip for the waiter, and drove back home.

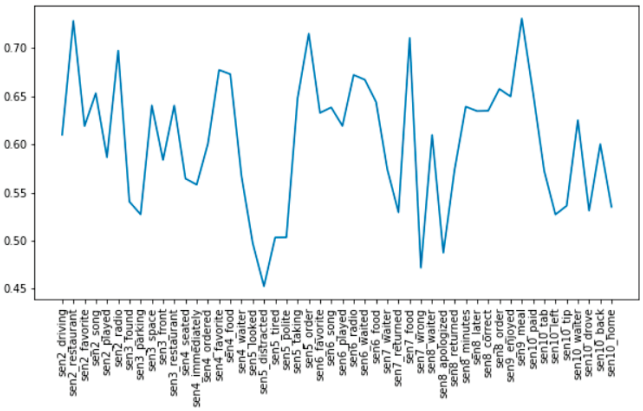

Figure 1: Moving cosine similarity chart of the sample story seen in Table 1.

with a fixed threshold. To this end, we gradually reduce this threshold until (and if) we find a set of outlier words that belong to exactly two sentences in the story.

We attempt to calculate a quantitative metric M that reflects the level of anomaly for the words involved in a possible case of foreshadowing. This measure will then be an estimation of how much predictive inference we believe a case of foreshadowing causes. If we do not find any such 2 anomalous sentences, then $M = 0$. Otherwise, we first calculate A and B as follows.

- A is the cosine distance between the mean of all of the word vectors in the story, and the mean of the set of outlier words:

$$A = 1 - cosine(mean_{all\_words}, mean_{outliers})$$

- B is the mean of the cosine similarity of each word in the outlier set with the mean of all others in that set:

$$mean(cosine(w_i, mean(\{outliers\} - w_i))$$

for every w_i in outliers.

A represents a measure of how anomalous the hint and twist in the foreshadowing are, by calculating the cosine distance between their means. B represents how contextually and semantically related the sets of outlier words (separated into two sentences) can be considered.

In order to yield M as a singular measure, we sum A with the absolute value of B. The reason for taking the absolute value for B is that we want to consider the semantic relationship and not necessarily similarity. If two sets of outlier words across two sentences have a cosine similarity of -1 (semantically opposite), that is potentially just as valuable for providing a hint as $B = 1$. In order to limit M to $[0, 1]$, we note that the range of A is $[0, 2]$ and B's is $[0, 1]$. Thus:

$$M = \frac{A + abs(B)}{3}$$

It is important to note that M is not a probability, and hence, a value of 1 (while highly unlikely) would not mean anything special. For instance, these values for our example story with a clear case of foreshadowing are $A = .35$, $B = .17$ and $M = .40$.

Our proxy measure of predictive inference and foreshadowing has a two-fold output. Firstly, it has an output of being zero or non-zero. This binary output shape is driven by the nature of foreshadowing, which necessarily consists of two places in the story that have semantic links between them. Secondly, once the method does find a pair of candidate sentences for foreshadowing, it then estimates how much predictive inference it may cause. Many forms of foreshadowing involve subtle hints that do not necessarily cause outlier word vectors. Predictive inference, however, is caused when the audience notice a form of discrepancy, inconsistency or curious detail in their perception of the sequence of events. Based on these intuitions, it is plausible to imagine that the kinds of foreshadowing cases that are capable of driving predictive inference are also more likely to involve outlier word vectors.

Evaluation

In order to evaluate our approach, we conducted a two-phased user study.

Study Phase I

In the first phase, we used 3 short and simple stories about going to a restaurant, going on a plane flight and a bank robbery ("restaurant", "flight", and "bank robbery" stories accordingly). These stories were extracted from (Li et al. 2013) as largely mundane event sequences that lacked specificities. Each story contained 10 to 14 short sentences and all 3 stories yielded an $M = 0$ with our method.

We recruited 40 participants on Mechanical Turk and asked every participant to add a "HINT" and a "TWIST" to each of the 3 stories. Participants were asked to specify the locations in the story where their HINT and TWIST would

be added (between any two sentences or after the last one), and were given an open-ended text field to write their additions. While no length limit was enforced, participants were encouraged to limit their HINT and TWIST additions to 1 sentence each. They were not able to change the existing sentences in the stories. This evaluation was intended to find how reliable our proxy measure is in finding cases of foreshadowing.

Results. After cleaning the data (removing 4 participants' data who entered random words), this step resulted in a dataset of 108 stories with foreshadowing (36 for each story). We ran our method on all of these stories to find out the percentage of them for which our proxy measure yields a value of $M > 0$. Table 2 shows this ratio for each of the 3 original stories and overall, as well as the mean M values.

Table 2: Phase 1 results. Ratio denotes the percentage of the stories with foreshadowing for which our proxy measure results in an $M > 0$. The mean M value for each original story is denoted as $mean(M)$. This average is calculated only for non-zero M values.

Story	**Ratio**	$mean(M)$
restaurant	78%	.34
flight	75%	.30
bank robbery	94%	.29
Overall	83%	.31

Study Phase II

In the second phase, we investigated the links between M and the perceived interestingness of stories. For each of the 3 stories, we picked 2 random instances from the output of the first study: one with a high M value (randomly selected from the top 5) and one with a non-zero low M value (randomly selected from the bottom 5, excluding the ones with $M = 0$). We did not choose the stories with $M = 0$ since those are clearly missed by our proxy measure, and hence might or might not drive a high level of cognitive interest. This evaluation sought to investigate the differences in the human perception of the stories with high and low M values.

We recruited 52 participants from Mechanical Turk (different than the participants of the phase 1), and in a within-subject design, asked them to rate the interestingness of the 6 stories selected above on a Likert scale (1-5) and in a randomized order. This resulted in 52 ratings for for each of the 6 selected stories.

Results. Table 3 shows the mean and median rating of each of the 6 selected stories. We used a one-tailed Wilcoxon Signed-Rank test to look for statistically significance differences between the ratings of the two versions of each story (High-M and Low-M).

Discussions

The first phase's results, shown in Table 2, indicate that our proxy measure performs well with a rate of $> 75\%$, across the three different original stories and cases of foreshadowing authored by 36 participants. We did not see a major concentration on a sub-group of participants for the stories with undetected foreshadowing. However, out of 19 such undetected cases, 8 of them belonged to 4 users (2 cases each). This observation can speak to the impact of individual style of writing in foreshadowing or the level of subtlety of the hints. Foreshadowing can involve long causal chains or contextual semantics links that do not depend on words that co-occur in other contexts (and hence their word vectors do not yield high cosine similarities if trained on general corpora).

Table 3: Phase 2 results showing the means and medians of the Likert scale ratings of the 6 selected stories (one High-M and one Low-M sample story for each of the 3 main story categories (restaurant, flight, bank robbery)). The p-value is from a one-tailed Wilcoxon Signed-Rank test.

Story	**Mean**	**Median**	**p-value**
restaurant, High-M	2.7	3	**.032**
restaurant, Low-M	2.3	2	
flight, High-M	2.7	3	.085
flight, Low-M	2.9	3	
bank robbery, High-M	3.81	4	**.038**
bank robbery, Low-M	3.58	4	

As previously mentioned, the two-fold output of our proxy measure also allows us to estimate how much predictive inference a case of foreshadowing makes. Predictive inference is likely affected by many factors, including the more subjective "emotional" interests discussed before. Thus, our proxy measure's estimation is mainly based on the intuition that if a set of outlier words are semantically farther away from what the rest of the story has been about, they are more likely to raise a question mark for the audience and drive predictive inference. In simple terms, the farther such distance is, the bigger the mental question mark of the reader can be. For all 3 of our original stories, the average M listed in Table 3 is about .30, with a maximum value of .42, .40, .37, minimum of .25, .18, .18, and a standard deviation of .04, .06 and .06, respectively. These results indicate that our proxy measure has some level of variation, but the variance is small enough that one can categorize the level of estimated predictive inference in "high" and "low" groups. It is plausible that for other datasets or longer stories this variance could grow.

Consistent with such categorization, the second phase of the study found statistically significant differences in the perceived interestingness of randomly chosen stories with high and low M amounts for 2 of the 3 original stories (restaurant and bank robbery). The samples of the other story (flight) with high and low M amounts did not show a statistical significance in the difference of their perceived interestingness by our participants.

It is noteworthy that the bank robbery story shows higher perceived interest levels than the other two stories, as a plot that involves *danger*, one of the "absolute interests" (Schank 1979) introduced earlier in this paper. Future direction of research can investigate the effects of cognitive interests

such as predictive inference on all "experiential interests" (Behrooz et al. 2018) such as absolute interests, emotional (Kintsch 1980) or topic interests (Campion 2004).

Moreover, Campion empirically shows that predictive inference is regarded as "hypothetical facts" by the audience (Campion 2004) (versus the deductive inferences that are regarded as certain facts). It would be interesting to investigate ways to affect a controllable process of neural story generation (Tambwekar et al. 2018b) based on such hypothetical facts in order to have more complicated structures that lead to predictive inference.

Regardless of the method used to generate the narrative, a situated use-case of storytelling for interactive agents involves a selection of events from agent's memory that are likely to be interesting, such as in (Behrooz, Swanson, and Jhala 2015). Importantly, the agent would then have to choose which details to include in the story, which specifications of the elements or actions to add or in what order to say them (Montfort 2009). Having a proxy measure for cognitive and other kinds of interest can guide such selection so that the agent would choose the details and specifications that may cause the most interest.

Conclusion

Proxy measures that can evaluate the quality of generated stories are essential to the enhancement of the generative methods and models (Purdy et al. 2018). While linguistic features and semantic coherence are essential elements of a good story, the real key to a successful story-based interaction is the perceived *interestingness*. Assessing the perception of the story is a complicated task; it can be very subjective, contextual, and hard to measure. However, there are theories of story interestingness that we can rely on. In particular, "cognitive interest" tends to be less subjective or contextual and mainly a product of the cognitive processes involved in the perception. We specifically target predictive inference as a root cause of cognitive interest (Campion, Martins, and Wilhelm 2009). Our method uses contextual word embeddings (BERT) to find cases of foreshadowing in a given short story, as a common cause of predictive inference. In a study, we show that this method can find a majority of cases of foreshadowing authored by our participants. Moreover, our proxy measure associates a value to the level of predictive inference that a story is likely to cause. Our study found significant differences in the perceived interestingness of stories with low and high such value.

References

Admoni, H., and Scassellati, B. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6(1):25–63.

Bae, B.-C., and Young, R. M. 2008. A use of flashback and foreshadowing for surprise arousal in narrative using a plan-based approach. In *Joint International Conference on Interactive Digital Storytelling*, 156–167. Springer.

Battaglino, C., and Bickmore, T. 2015. Increasing the engagement of conversational agents through co-constructed storytelling. In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*.

Behrooz, M.; Mobramaein, A.; Jhala, A.; and Whitehead, J. 2018. Cognitive and experiential interestingness in abstract visual narrative. *Cognitive Science Society (CogSci)*.

Behrooz, M.; Swanson, R.; and Jhala, A. 2015. Remember that time? telling interesting stories from past interactions. In *Interactive Storytelling*. Springer.

Bickmore, T., and Cassell, J. 1999. Small talk and conversational storytelling in embodied conversational interface agents. In *AAAI fall symposium on narrative intelligence*, 87–92.

Campion, N.; Martins, D.; and Wilhelm, A. 2009. Contradictions and predictions: Two sources of uncertainty that raise the cognitive interest of readers. *Discourse Processes* 46(4):341–368.

Campion, N. 2004. Predictive inferences are represented as hypothetical facts. *Journal of Memory and Language* 50(2):149–164.

Chen, D. L., and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 190–200. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long Papers)*, 2250–2260.

Coon, W.; Rich, C.; and Sidner, C. L. 2013. Activity planning for long-term relationships. In *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013, Proceedings*, volume 8108, 425. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elson, D. K. 2012. *Modeling narrative discourse*.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Farrell, R., and Ware, S. G. 2016. Fast and diverse narrative planning through novelty pruning. In *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Foltz, P. W.; Kintsch, W.; and Landauer, T. K. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes* 25(2-3):285–307.

Freebody, P., and Anderson, R. C. 1986. Serial position and rated importance in the recall of text. *Discourse processes* 9(1):31–36.

Frick, R. W. 1992. Interestingness. *British Journal of Psychology* 83(1):113–128.

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

Gladkova, A.; Drozd, A.; and Matsuoka, S. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, 8–15.

Goldberg, Y., and Levy, O. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Gomes, P.; Paiva, A.; Martinho, C.; and Jhala, A. 2013. Metrics for character believability in interactive narrative. In *International Conference on Interactive Digital Storytelling*, 223–228. Springer.

Graesser, A. C.; McNamara, D. S.; Louwerse, M. M.; and Cai, Z. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36(2):193–202.

Hidi, S., and Baird, W. 1986. Interestingness—a neglected variable in discourse processing. *Cognitive Science* 10(2):179–194.

Iran-Nejad, A. 1987. Cognitive and affective causes of interest and liking. *Journal of Educational Psychology* 79(2):120.

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

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Khalifa, A.; Barros, G. A.; and Togelius, J. 2017. Deeptingle. *arXiv preprint arXiv:1705.03557*.

Kim, S.-i. 1999. Causal bridging inference: A cause of story interestingness. *British Journal of Psychology* 90(1):57–71.

Kintsch, W. 1980. Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics* 9(1-3):87–98.

Lebowitz, M. 1987. Planning stories. In *Proceedings of the 9th annual conference of the cognitive science society*, 234–242.

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*.

Mandler, G. 1982. The structure of value: Accounting for taste. *Center for Human Information Processing Report* 101.

Martin, L. J.; Ammanabrolu, P.; Wang, X.; Hancock, W.; Singh, S.; Harrison, B.; and Riedl, M. O. 2017. Event representations for automated story generation with deep neural nets. *arXiv preprint arXiv:1706.01331*.

McCoy, J.; Treanor, M.; Samuel, B.; Mateas, M.; and Wardrip-Fruin, N. 2011. Prom week: social physics as gameplay. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, 319–321. ACM.

Meehan, J. R. 1977. Tale-spin, an interactive program that writes stories. In *IJCAI*, volume 77, 91–98.

Montfort, N. 2009. Curveship: An interactive fiction system for interactive narrating. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 55–62. Association for Computational Linguistics.

Ontañón, S., and Zhu, J. 2010. Story and text generation through computational analogy in the riu system. In *Sixth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

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.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

PÉrez, R. P. Ý., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Porteous, J.; Cavazza, M.; and Charles, F. 2010. Applying planning to interactive storytelling: Narrative control using state constraints. *ACM Transactions on Intelligent Systems and Technology (TIST)* 1(2):10.

Purdy, C.; Wang, X.; He, L.; and Riedl, M. 2018. Predicting generated story quality with quantitative measures.

Riedl, M. O., and Young, R. M. 2005. An objective character believability evaluation procedure for multi-agent story generation systems. In *Intelligent Virtual Agents*, 278–291.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39(1):217–268.

Schank, R. C. 1979. Interestingness: controlling inferences. *Artificial intelligence* 12(3):273–297.

Sidner, C. L. 2015. Engagement, emotions, and relationships: On building intelligent agents. *Emotions, Technology, Design, and Learning* 273.

Swanson, R., and Gordon, A. S. 2008. Say anything: A massively collaborative open domain story writing companion. In *Interactive Storytelling*. Springer. 32–40.

Tambwekar, P.; Dhuliawala, M.; Mehta, A.; Martin, L. J.; Harrison, B.; and Riedl, M. O. 2018a. Controllable neural story generation via reinforcement learning. *arXiv preprint arXiv:1809.10736*.

Tambwekar, P.; Dhuliawala, M.; Mehta, A.; Martin, L. J.; Harrison, B.; and Riedl, M. O. 2018b. Controllable neural story generation via reinforcement learning. *arXiv preprint arXiv:1809.10736*.

Turner, S. R. 1994. Minstrel: A computer model of creativity and storytelling.

Turner, S. R. 2014. *The creative process: A computer model of storytelling and creativity*. Psychology Press.

Wilensky, R. 1983. Story grammars versus story points. *Behavioral and Brain Sciences* 6(04):579–591.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, 19–27.