Fabula Entropy Indexing: Objective Measures of Story Coherence

Louis Castricato, Spencer Frazier, Jonathan Balloch, and Mark O. Riedl
Georgia Tech
{lcastric, sfrazier7, balloch}@gatech.edu, {riedl}@cc.gatech.edu

arXiv:2104.07472v1 [cs.CL] 23 Mar 2021

Abstract

Automated story generation remains a difficult area of research because it lacks strong objective measures. Generated stories may be linguistically sound, but in many cases suffer poor narrative coherence required for a compelling, logically-sound story. To address this, we present Fabula Entropy Indexing (FEI), an evaluation method to assess story coherence by measuring the degree to which human participants agree with each other when answering true/false questions about stories. We devise two theoretically grounded measures of reader question-answering entropy, the entropy of world coherence (EWC), and the entropy of transitional coherence (ETC), focusing on global and local coherence, respectively. We evaluate these metrics by testing them on human-written stories and comparing against the same stories that have been corrupted to introduce incoherencies. We show that in these controlled studies, our entropy indices provide a reliable objective measure of story coherence.

1 Introduction

Automated story generation is one of the grand challenges of generative artificial intelligence. AI storytelling is a crucial component of the human experience. Humans have always used storytelling to entertain, share experiences, educate, and to facilitate social bonding. For an intelligent system to be unable to generate a coherent story limits its ability to interact with humans in naturalistic ways.

There have been a number of techniques explored for story generation; these include symbolic planning, case-based reasoning, neural language models and others. Despite extensive research, automated story generation remains a difficult task.

One of the reasons why automated story generation is such a difficult area of research is due to weak objective validation measures. Traditional automated measures of natural language quality—perplexity and n-gram based methods such as BLEU (Papineni et al., 2002)—are insufficient in creative generation domains such as story generation. These metrics assume that generated language can only be good if is resembles testing data or a given target story. This precludes the possibility that stories may be good yet be completely novel. Indeed, the goal of story generation is usually the construction of novel stories.

In the absence of automated evaluation metrics, the alternative is to use human participant studies. Human participants, typically recruited via crowdsourcing platforms (e.g Mechanical Turk or Prolific), are asked to read the stories generated by various systems and provide subjective rating or rankings. Questionnaires may ask participants to rate or rank the overall quality of stories, but may also ask specific questions about features of stories such as fluency or coherence. Coherence is particularly difficult feature of stories to measure because the term "coherence" can mean different things to different participants.

In this paper, we introduce a technique for **objective** human participant evaluation, called *Fabula Entropy Indexing* (FEI). FEI provides a structure for metrics that more objectively measure story coherence based on human question-answering. A *fabula* is a narratological term referring to the reader's inferred story world that a story takes place in, whether it be similar to the real world or a fantasy or science fiction world. The reader may of course be surprised by certain events but other events may seem implausible or contradictory, thus disrupting coherence. As they read, humans form cognitive structures to make sense of a story, which in turn can be used to answer simple true/false questions about the story. As such, an *incoherent* story results in readers making random guesses about the answers to these questions. FEI metrics thus measure the entropy of the answers—how much the answers disagree with each other—which directly

correlates with the coherence of the story.

We introduce two such FEI metrics: *Entropy of Transitional Coherence* (ETC) and *Entropy of World Coherence* (EWC), measuring (respectively) sequential coherence between events in a story, and the internal coherence of the *story world*: the facts about characters, objects, and locations that distinguish a story. The correlation between human question-answering and these metrics are grounded in narratological[1] theories.

To validate the measure, we test our metrics on human-written stories as well as corrupted versions of those stories. For the corrupted stories, we artificially reduce the coherence by altering elements of the story. We show that FEI metrics evaluate non-corrupted human-written stories as having low entropy and corrupted stories as having higher entropy.

2 Background and Related Work

2.1 Automated Story Generation

Early story and plot generation systems relied on symbolic planning (Meehan, 1976; Lebowitz, 1987; Cavazza et al., 2003; Porteous and Cavazza, 2009; Riedl and Young, 2010; Ware and Young, 2011) or case-based reasoning (Pérez y Pérez and Sharples, 2001; Peinado and Gervás, 2005; Turner, 2014). An increasingly common machine learning approach to story generation is to use neural language models (Roemmele, 2016; Khalifa et al., 2017; Clark et al., 2018; Martin et al., 2018). These techniques have improved with the adoption of Transformer-based models, such as GPT-2 (Radford et al., 2019). While GPT-2 and similar neural language models are considered highly fluent from a grammatical standpoint.

In these systems, a neural language model learns to approximate the distribution $P_\theta(tok_n|tok_{<n})$ where θ is the parameters that approximate the pattern of an underlying dataset. Stories are produced by providing an initial context sequence, then iteratively generating additional tokens by sampling from the distribution. When the language model is trained on a corpus of stories, subsets of the generated text tend to also be a story.

One of the reasons why story generation is challenging is because of the strong requirement that stories be *coherent*. Coherence can refer to readability/fluency. However, stories also require *plot coherence*, which is how well the elements of a plot cohere with each other. Studies of human reading comprehension (Trabasso and Van Den Broek, 1985; Graesser et al., 1991, 1994) show that humans comprehend stories by tracking the relations between events. Reader comprehension studies suggest that readers rely on the tracking of at least four types of relations between events: (1) causal consequence, (2) goal hierarchies, (3) goal initiation, and (4) character intentions. The perceived coherence of a story is a function of the reader being able to comprehend how events correlate to each other causally or how they follow characters' pursuits of implicit goals.

To control the generation and achieve greater coherence, a high-level plot outline can either be generated or given as an input to a language model. (Fan et al., 2018; Peng et al., 2018; Rashkin et al., 2020; Brahman and Chaturvedi, 2020). These techniques can produce more coherent stories when their guidance forces different parts of the story to appear related or to follow a pattern acceptable to humans.

Tambwekar et al. (2018) attempt to train a neural language model to perform goal-based generation. They fine-tune a neural language model with a policy-gradient reinforcement learning technique that rewards the language model for generating events progressively closer to the goal event.

2.2 Story Generator Evaluation

Traditional automated measures of natural language quality such as perplexity or n-gram comparisons (e.g., BLEU) are generally considered insufficient for evaluating story generation systems. Perplexity is the measure of how well a model captures the patterns in an underlying dataset. Implicit in the notion of perplexity is the belief that the quality of a model is tied to its ability to reconstruct its own data. However, in automated story generation, stories that are very dissimilar to training and testing data can also be "good". Likewise, BLEU (and related techniques such as ROGUE and sentence mover techniques (Clark et al., 2019)) measure a language model's ability to produce n-grams in a specific target sentence, whereas a good story may not resemble a given target story and yet still be coherent.

The gold standard for evaluation of automated story generation systems is to use human participant studies. Many systems are evaluated with subjective questionnaires in which human partic-

[1] *Narratology* is the study of stories and storytelling.

ipants either rate generated stories on a scale, or rank pairs of stories. Often a single question is asked about overall quality. Other subjective questions focusing on different story attributes, such as coherence, may be asked as well. Asking questions about coherence is tricky as participants may have different notions of what coherence might mean, from grammatical notions of coherence to logical story structure.

Purdy et al. (2018) introduced a set of subjective questions for human participant studies about global coherence, local consistency, grammaticality, and overall story quality. Algorithms to predict how humans would answer these questions were also introduced. The goal of this work was to reduce reliance on expensive human-participant studies. One innovation is that they don't directly ask about coherence, which can be an ambiguous term, but instead ask questions such as "the story appears to be a single plot". This set of questions has been used by Tambwekar et al. (2019) and Ammanabrolu et al. (2020). The algorithms introduced by Purdy et al. (2018) were validated and proven to be reliable predictors but the measure of coherence was shown to be the weakest predictor.

The USER technique, introduced as part of Storium (Akoury et al., 2020), is a means of evaluating stories by giving human participants the means to edit a generated story. They measure the largest subsequence not edited by the author during a story continuation. They conclude that their measure is strongly correlated with human evaluation of coherency.

Li et al. (2013) evaluated their story generation system using an objective human participant study. They generated stories and then had human add sentences, delete sentences, or swap sentence orderings. The number of edits is used to score the story generation system (lower is better).

Riedl and Young (2010) also evaluated their story generation system with an objective human participant study based on cognitive science. They conducted a question-answering protocol to elicit the cognitive model that humans had about the causal relations and goals of characters. Specifically they constructed a number of questions that the story generation system believed human readers should be able to answer. The measure of story quality was the degree to which humans answered the questions the way the algorithm predicted they would. This technique is the most similar in nature to our proposed measure of coherence; our technique is mathematically grounded and not tied to any particular way of generating stories.

3 Preliminaries

In this section we review narratological definitions that will be relevant to understanding how to measure the Fabula Entropy Indices.

Definition 3.1. A **narrative** is the recounting of a sequence of events that have a continuant subject and constitute a whole (Prince, 2003).

An event describes some change in the state of the world. A "continuant subject" means there is some relationship between the events—it is about something and not a random list of unrelated events. All stories are narratives, but also include some additional criteria that are universally agreed upon.

Structural narratologists suggest there are different layers at which narratives can be analyzed: *fabula* and *syuzhet* (Bal and Van Boheemen, 2009)

Definition 3.2. The **fabula** of a narrative is an enumeration of all the events that take place the story world.

Definition 3.3. The **syuzhet** of a narrative is a subset of the fabula that is presented via narration to the audience.

The events in the fabula are temporally sequenced in the order that they occur, which may be different than the order in which they are told. Most notably, the events and facts in the fabula might not all exist in the final telling of the narrative; some events and facts might need to be inferred from what is actually told. It is not required that the syuzhet to be told in chronological order, allowing for achronological tellings such as flash forward, flashback, ellipses (gaps in time), etc.

They key is that readers interact more closely with syuzhet and must infer the fabula through the text of the syuzhet. Because a fabula inferred, it may be occuring in one of many possible worlds in a modal logic sense (Ryan, 1991).

Definition 3.4. A **story world** is a set of possible worlds that are consistent with the facts and events presented to the reader in the syuzhet.

As events and facts are presented throughout the narrative, the probability cloud over story worlds collapses and a reader's beliefs become more certain.

Events in the fabula and story world have different degrees of importance:

Definition 3.5. A **kernel** is a narrative event such that after its completion, the beliefs a reader holds as they pertain to the story have drastically changed.

Definition 3.6. A **satellite** is a narrative event that supports a kernel. They are the minor plot points that lead up to major plot points. They do not result in massive shift in beliefs.

Satellites imply the existence of kernels, e.g. small plot points will explain and lead up to a large plot point, but kernels do not imply the existence of satellites—kernels do not require satellites to exist. A set of satellites, $s = \{s_1, \ldots, s_n\}$, is said to be relevant to a kernel k if, after the kernel's competition, the reader believes that the set of questions posed by k are relevant to their understanding of the story world given prior s.

An implication of kernels and satellites is that one can track a reader's understanding of a story over time by asking the reader questions relevant to the story before and after each major plot point. As kernels change the reader's beliefs about the story world and the fabula, then their answers to questions change as well.

4 Fabula Entropy Indexing

Fabula Entropy Indexing (FEI) measures story coherence based on human question-answering. Humans build cognitive structures to make sense of a story, which in turn can be used to answer simple true/false questions about the story. A coherent narrative results in readers having well-formed cognitive models of the fabula and story world(Graesser et al., 2003; Trabasso et al., 1982). Because the cognitive models formed during reading are predictable across readers one can infer that coherent stories result in readers being more likely to answer questions about a story similarly (Graesser et al., 1991). *Incoherent* stories thus result in readers making random guesses about the answers to questions. FEI looks at the entropy of the answers—how much readers disagree with each other—as a signal of coherence of the story.

We decompose FEI into two separate metrics. *Entropy of Transitional Coherence* (ETC) measures the necessity of transitional ordering: in time t, event or fact x is necessary to maintain a story's coherence. In other words, was this fact probable before t? This establishes whether a reader could reasonably anticipate the occurring between two events. *Entropy of World Coherence* (EWC) on the other hand is not time dependent. EWC measures the probability of an event or fact y occurring *at any time* in a story world.

The core idea of Fabula Entropy Indexing is that readers can be asked true/false questions and that the agreement in readers' answers indicates coherence. However, questions must take the form of implications $q : A \implies B$ (read "if A then B") and the two propositions A and B must have *relevance* to each other.

Definition 4.1. For a question about a story, q, of the form "if A then B" with possible values for $A = \{T, F\}$ and possible values for $B = \{T, F\}$. Identifying A with the set of possible answers to it, we say that the **relevance** of B to A given some prior γ is

$$H(A = a_i|\gamma) - H(B = b_j|A = a_i, \gamma) \quad (1)$$

where a_i and b_j are the true answers to A and B and H refers to binary entropy. (Knuth, 2004).

Note that the relevance of B to A *depends on the ground truth*. Consider the case where A is "is Harry Potter the prophesied Heir of Slytherin?" and B is "can Harry Potter speak Parseltongue because he is a descendent of Slytherin?". If Harry is a blood descendant of Slytherin and that is why he can speak Parseltongue, then B is highly relevant to A. However, the actual truth of the matter is that Harry's abilities are completely independent of his heritage. Therefore B does not have relevance to A even though *it could have had relevance* to A had the ground truth been different.

4.1 Entropy of Transitional Coherence

Certain facts or events in stories have temporal dependencies. For example, a protagonist may hammer a nail into the wall. If subsequent events reveal the fact that the protagonist never held a hammer this causes temporal or transitional incoherence.

If we force our question to be an implication, namely of the form "Given that A occurs within the story, then B", we are attempting to determine the relevance of a query B to a query $A = true$, specifically:

$$H(A = true|\gamma) - H(B = b_j|A = true, \gamma).$$

If A is given within the reader's inferred fabula, then A is always true and we simply want to query about B. However if A is undetermined within the reader's inferred fabula then we are as a whole

querying about "If A then B," and forcing the reader to reconcile both A and B without any belief about A.

Entropy of Transitional Coherence therefore asks questions of readers in which A is a belief from before a kernel and B is a belief from after a kernel. Let question q be of the form "Given that A occurs within the story, then B." That is $q := A \implies B$. Let $P(q)$ refer to the proportion of story worlds where q is true. The stronger the reader's belief, the more possible worlds in which q is true, and the higher the probability. Across all readers answering the question:

$$\begin{aligned} H(P(q)) &= H(q|\gamma) \\ &= H(A = T|\gamma) - H(B = b_j|A = T, \gamma) \end{aligned} \tag{2}$$

By averaging across all questions Q that span kernels, we arrive at the definition of ETC:

$$E(Q) = \frac{1}{|Q|} \sum_{q \in Q} H\big(P(q)\big) \tag{3}$$

In the context of Entropy of Transitional Coherence, $ETC(Q) = E(Q)$.

Consider the following example for discussing the importance of ETC. A person needed a bath, so they went for a run. A possible query here would be "Given a person needed a bath, does this contradict that they went for a run?" In this particular example, we can assume going for a run is a kernel and as such this query measures if needing a bath is a plausible precondition to desiring to go on a run. Equivalently, does the reader believe "If the person needs a bath, then they go for a run." If the story makes less sense to the reader, the reader attempts to reconcile these two clauses and as such would be more likely to guess.(Trabasso et al., 1982; Mandler and Johnson, 1977)

4.2 Entropy of World Coherence

Whereas Entropy of Transitional Coherence measures coherence as events cause the story world to change, Entropy of World Coherence (EWC) measures the the coherence of static fact about the story world. For example if a story contains a protagonist that is described as being short but is also described as hitting their head on the top of a doorframe, we might find readers have more varied responses to a question about the protagonist's height.

Entropy of World Coherence also uses Equation 3 (that is, $EWC(Q) = E(Q)$) but does not require that the questions reference before and after kernels. There need not be any temporal requirement to questions. Instead EWC relies on questions about descriptive elements in a story, as signified by adjective and adverbs. However, these descriptions of characters, objects, or places must be integral to at least one event in the narrative.

4.3 Measuring Coherence with Human Participant Studies

Having mathematically defined our two coherence metrics, ETC and EWC, as a function of readers responding to a set of questions about temporal or non-temporal aspects of a story, we now describe how we use ETC and EWC to measure coherence of stories, particularly those from by automated story generation systems. There are three key steps to Fabula Entropy Indexing as a methodology.

The first step is to use an automated story generation system to generate a number of stories that are representative of its capabilities. Typically this would be done by randomly seeding the generator.

The second step is to produce a number of questions. To produce questions for ETC, one identifies the kernels—the major plot points—and constructs questions such as:

- Does Entity A's sentiment/emotion change between line N-1 and N?
- Does Object A change possession in Line N+1?

To produce questions for EWC, one identifies adjectives and adverbs that could be changed, such as:

- Does [Adverb/Adjective] contradict an assertion on Line N?
- Could [Adverb/Adjective] be removed and the story world would remain unchanged?

One would want to produce as many questions as possible. Note that while the questions above do not read as implications immediately, they can be expressed as the required implications after a bit of work and thus still satisfy our constraint.

It doesn't matter what the questions are or what the answers are—we do not require a ground truth—as long as the questions reference aspects of the story that can impact readers' cognitive model formation. ETC and EWC guide us toward kernels and attributes, respectively. Fabula Entropy Indexing

measures coherence by observing the agreement between human participants when answering these questions.

The third step is to recruit human study participants to read a story and then answer the associated questions. There is no ground-truth "correct" answers—we are not testing participants ability to answer in a certain way. Instead, we use Equation 3 to measure agreement between responses, under the assumption that more coherent stories prompt readers to construct more consistent mental models of the fabula and story world.

ETC and EWC can be compared between representative sets of stories between different automated story generation systems. Lower entropy values implies greater coherence.

5 Experiments

To validate Fabula Entropy Indexing in general, and ETC and EWC in particular, we need to verify that the methodology in Section 4.3 produces low entropy values for coherent stories and high entropy values for incoherent stories. Because automated story generation is still an open research question, we validate ETC and EWC on human-written stories that are known to be coherent. We assume that human-written stories are coherent. To compare entropy indices against incoherent stories, we devise a technique for corrupting human written stories in particular ways that are likely to result in incoherent stories. Exemplar corruptions include negating adjectives, swapping events from different stories or randomly changing key descriptors of characters.

5.1 Entropy of World Coherence Stories

For EWC, we source a number of short stories by authors such as Rumi, Tolstoy and Gibran. Specifically, this is a subset available in a public repository[2] unaffiliated with the authors of this paper. For each story we subdivide them into 10-line segments if the story was longer than 10 lines. We selected 9 stories for the experiment.[3]

To create a corrupted story baseline in which story coherence is less assured, we copied the 9 stories and made changes to them. We recruited 4 participants who are unaffiliated with the research team and asked them to independently select a subset of the adjectives and adverbs from a story and swap them for their antonyms. This produced stories that are, at a story world level, less coherent since due to the highly descriptive nature of the stories one swap was more likely to lead to a contradiction later on in the story. Participants were required to create the inconsistency and not to fix their incoherency with more swaps. Participants were compensated $20/hr to complete this task.

5.2 Entropy of Transitional Coherence Stories

For Transitional Coherence we require a direct correspondence between events and sentences. *Plotto* (Cook, 2011) is a compilation of plot points with annotations about which plot points can be followed by others. Plotto can thus be used to generate plot outlines assembled from human-written segments. The Plotto plot points contain few adjectives and plot outlines generated from the Plotto technique are unambiguous with respect to transitions in the story world. Since plotto consists of plot points, every vertex, and in our case line number, using the Plotto technique is a kernel. Within every kernel are a number of sentences, typically 2-3, that denote the satellites.

Since Plotto directly states plot points rather than having the reader infer them, this allows us to controllable corrupt the order of plot points by swapping lines- something that is rarely possible with human written short stories.

To construct stories for measuring ETC, we use the Plotto technique to generate 5-6 sentence short stories. For the experiment we generated 9 stories in this way.

To construct corrupted stories, we copied the 9 stories above and then swap the order of plot points, which results in incoherence (e.g. a burglar getting away with a crime before they're even born). We generate Plotto stories with 5 vertices, and randomly choose a span of 3 vertices. Within that span, we shuffle their order.

5.3 Question Generation

To measure ETC and EWC we require a set of true/false questions for each story. To ensure that we do not introduce experimental bias in questions for each story, we recruited 4 people to write questions for each story. Question writers were compensated $20/hr and produced 10-15 questions per

[2] https://github.com/pelagia/short-stories

[3] In both the ETC and EWC cases we had intended to evaluate over 10 stories but one story was rejected due to one of the stories inadvertently having a controversial interpretation when corrupted and which was only pointed out to us by one of the question-answering participants.

story.

For the corrupted sets of both Plotto and non-Plotto stories, we task a human participant to write questions guided by a set of templates which provide the best coverage over the more likely reader possible worlds. That is to say, if there were N reasonable interpretations of the story, we aimed to have our human subjects construct questions that could differentiate between N interpretations. Said another way, all templates probe the probability or plausibility of one plot point occurring or impacting the reader's comprehension of other plot points, in some way.

Participants were provided a packet which includes a description of the research, instructions for the task and a list of templates to follow when generating questions. Templates were also used to standardize the format of questions human participants in the subsequent experiment would receive. Question writing participants could freely choose the entities, properties and line numbers represented in each question.

A partial list of corruption prompts and a full list of question templates with some exemplar completions are provided in the Appendix.

5.4 Methodology

For each task, we recruit 180 participants on the Prolific platform, split evenly between ETC and EWC tasks. Demographic screening excluded any non-US, non-ESL individuals as well as those with linguistic impediments on the basis of the tasks' relative comprehension complexity. Each worker was either given corrupted stories or uncorrupted stories, but never both. This was done to prevent a worker from seeing both the uncorrupted and corrupted version of a story and as such biasing the results. Every worker received a randomized set of 3 stories. For each story, 10-15 yes or no questions were asked about interdependencies between sentences of the same story. Workers were compensated $20/hr for their time and given a screening question that was a handmade EWC and ETC example respectively. These examples were not used in computing the final result.

5.5 Results

The results are summarized in Figure 1 for Entropy of Transitional Coherence and Figure 2 for Entropy of World Coherence. The bars on the left are the results for uncorrupted, original stories and the bars on the right are for the stories modified to corrupt

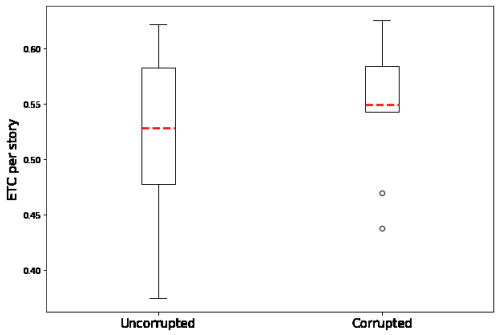

Figure 1: Entropic indices of transitional coherence derived from human participant evaluation of Plotto stories. Lower is better.

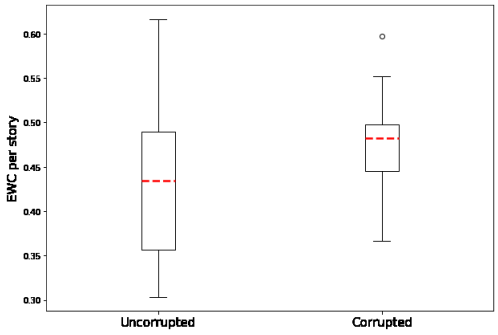

Figure 2: Entropic indicies of world coherence derived from human participant evaluation of the non-Plotto story dataset. Lower is better.

coherence. The red line indicates the mean of each distribution. Median is not reported. The results suggest that original stories have lower entropy and are thus more coherent. This validates fabula entropy indexing because the corruptions we applied to the same set of stories are designed to interfere with readers' abilities to form a well-formed model of the fabula and story world.

We do not report statistical significance because statistical significance tests are undefined on entropy distributions, which are not probability distributions.

6 Discussion

From the results, we can make some observations. The first is that the corrupted stories are not a traditional experimental baseline. The corruptions were designed to show that intentionally introduced incoherencies do in fact result in an increase in entropy. Second, the corruptions are designed to introduce the smallest possible amount of incoherence to stories as possible. Therefore, we would not expect a large increase in entropy due to a single corruption per story. The fact that entropy increases with

the introduction of minimalist corruptions indicates that Fabula Entropy Indexing is sensitive to such small changes. We would anticipate an automated story generator that routinely makes transitional or world coherence errors to result in much more significant differences in entropy values.

The entropies for corrupted stories have more dense distributions. Not only was there more disagreement about the answers to questions, but the disagreement was consistent across all stories. This is to be expected because the corruptions are synthetically designed to damage story coherence. The entropy distributions for real stories was spread over a wider range of entropy values per story.

ETC might not be as strong a metric as EWC. The average ETC of uncorrupted stories is higher than the EWC of uncorrupted stories. This may be due to (a) human tolerance for event ordering variations; (b) the Plotto technique may have produced plots in which plot points are only loosely connected; (c) our swap-based corruptions may not always produce incoherent stories.

The quality of the entropy indices are highly dependent on the extent to which the true/false questions target points in the story where potential incoherence can arise. It may theoretically be possible for some automated story generators to automatically generate good sets of questions, however this is currently an open research problem. The authors of this paper could have generated a better set of true/false questions targeting ETC and EWC than those unaffiliated with the research. However, doing so introduces the possibility of experimenter bias, which needs to be avoided by those who use this evaluation technique.

FEI has a couple of limitations. First, to measure ETC one must be able to identify kernels and make questions about elements before and after the kernels. Second, to measure EWC, the stories must be highly descriptive in nature and that there are plot points that are dependent on adjectives; many story generators do not produce descriptive texts.

FEI was validated on short stories, of 10 sentences or less. While there is no theoretical reason it will not work on longer stories, it will require substantially more questions to be produced and answered by human participant studies.

We have used the Fabula Entropy Indexing method described in this paper to evaluate an automated story generation system in (under review, 2021). The REDACTED system was designed explicitly to increase coherence of automatically generated stories over a large pretrained transformer language model baseline. The combined ETC and EWC for the experimental system were lower than the language model baseline. Moreover, we also compared the entropy indices of human-written baseline stories, showing that human stories result in lower entropy values than AI generated stories, which is to be expected at this time. This constitutes the first successful use of FEI for its intended purpose of evaluating automated story generation systems.

As part of the above real-world test case of FEI, we also performed a *subjective* human-participant study, showing that the entropy indices are low when humans report perceived coherence. We did not perform a subjective human participant study for this paper since we were working on stories that came from sources with reliable coherence.

7 Conclusions

Automated Story Generation research requires strong, reliable evaluation metrics, which have largely been absent, hampering research progress. We present the Fabula Entropy Indexing technique for objectively evaluating the coherence of stories. We demonstrate the effectiveness of this technique by showing how two FEI metrics, entropy world coherence and entropy transitional coherence, can be used to clearly discriminate between stories with and without coherence corruption. In contrast to subjective human participant studies, where it is challenging to get participants to answer questions about coherence, FEI provides a numerical rating of the coherence of stories that is grounded in theory.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. Storium: A dataset and evaluation platform for machine-in-the-loop story generation.

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering.

Mieke Bal and Christine Van Boheemen. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. *arXiv preprint arXiv:2010.06822*.

Marc Cavazza, Olivier Martin, Fred Charles, Steven J Mead, and Xavier Marichal. 2003. Interacting with virtual agents in mixed reality interactive storytelling. In *International Workshop on Intelligent Virtual Agents*, pages 231–235. Springer.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Elizabeth Clark, Yangfeng Ji, and Noah A Smith. 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)*, pages 2250–2260.

William Cook. 2011. *PLOTTO: the master book of all plots*. Tin House Books.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Art Graesser, Kathy L. Lang, and Richard M. Roberts. 1991. Question answering in the context of stories. *Journal of Experimental Psychology: General*, 120(3):254–277.

Art Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371–395.

Arthur C Graesser, Danielle S McNamara, and Max M Louwerse. 2003. What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking reading comprehension*, 82:98.

Ahmed Khalifa, Gabriella AB Barros, and Julian Togelius. 2017. Deeptingle. *arXiv preprint arXiv:1705.03557*.

Kevin H. Knuth. 2004. Measuring questions: Relevance and its relation to entropy. *AIP Conference Proceedings*.

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

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27.

Jean M Mandler and Nancy S Johnson. 1977. Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, 9(1):111–151.

Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

James Richard Meehan. 1976. The metanovel: writing stories by computer. Technical report, YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Federico Peinado and Pablo Gervás. 2005. Creativity issues in plot generation. In *Workshop on Computational Creativity, Working Notes, 19th International Joint Conference on AI*, pages 45–52.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49.

Rafael Pérez y Pérez and Mike Sharples. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(2):119–139.

Julie Porteous and Marc Cavazza. 2009. Controlling narrative generation with planning trajectories: the role of constraints. In *Joint International Conference on Interactive Digital Storytelling*, pages 234–245. Springer.

Gerald Prince. 2003. *A dictionary of narratology*. U of Nebraska Press.

Christopher Purdy, X. Wang, Larry He, and Mark O. Riedl. 2018. Predicting generated story quality with quantitative measures. In *AIIDE*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967*.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

Melissa Roemmele. 2016. Writing stories with help from recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Marie-Laure Ryan. 1991. *Possible worlds, artificial intelligence, and narrative theory*. Indiana University Press.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J Martin, Animesh Mehta, Brent Harrison, and Mark O Riedl. 2018. Controllable neural story plot generation via reinforcement learning. *arXiv preprint arXiv:1809.10736*.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J Martin, Animesh Mehta, Brent Harrison, and Mark O Riedl. 2019. Controllable neural story plot generation via reward shaping. In *IJCAI*, pages 5982–5988.

Tom Trabasso and Paul Van Den Broek. 1985. Causal thinking and the representation of narrative events. *Journal of memory and language*, 24(5):612–630.

Tom Trabasso et al. 1982. Causal cohesion and story coherence.

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

Stephen Ware and R Young. 2011. Cpocl: A narrative planner supporting conflict. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 6.

A Appendices

A.1 Alteration Templates[4]

The [Adjective1] Object/Entity/Event -> The [Adjective2] Object/Entity/Event

The [Adjective1] Object/Entity/Event -> The not [Adjective1] Object/Entity/Event

Object/Entity/Event is [Adverb1] [Adjective1] -> Object/Entity/Event is [Adverb1] [Adjective2]

Object/Entity/Event is [Adverb1] [Adjective1] -> Object/Entity/Event is [Adverb2] [Adjective1]

Object/Entity/Event [Adverb1][Verb] -> Object/Entity/Event [Adverb2][Verb]

These are just a small sample of templates given the complex nature of certain sentences. You can make alterations beyond this but adhere to the rules above.

A.2 Question Templates: EWC

In the context of this narrative setting, is [Adverb/Adjective] plausible? (e.g. an "otherworldly" dog showing up in a short story about World War 2 where you might otherwise describe a "stray" dog. Note: This may not be a constraint for all readers - those answering questions will only assess based on *their* belief about the world.)

Prior to this line did you imagine [Adverb/Adjective] was a possible descriptor for Object/Entity/Event?

After this line containing [Adverb/Adjective] do you hold the belief this is a possible descriptor or do you reject it?

Because of [Adverb/Adjective] does Line N contradict information in another line?

Because of [Adverb/Adjective] does this indicate emotional valence (extreme sentiment) toward an Object/Entity/Event?

In the line with [Adverb/Adjective] does this alter Author or Entity sentiment toward Object/Event?
Because of [Adverb/Adjective] does this change your sentiment toward some Entity/Object/Event?

Does [Adverb/Adjective] contradict an assertion on Line N?

Could [Adverb/Adjective] be removed and the story world would remain unchanged?

Without [Adverb/Adjective] on Line N, Line N+1 would not have happened.

A.3 Question Templates: ETC

Does Entity A's perception of Entity B change?

Do all Entities in Line N observe or gain awareness of Events in Line N+1?

Do the Events in Line N+1 contradict Events in Line N?

Does Entity A's sentiment/emotion change between line N-1 and N?

Does Object A still retain State S?

Does Object A change possession in Line

[4] Additional clarifying examples were given to participants when they requested them during task completion.

N+1?

Is Object A in Line N+1 necessary for Events in line N to occur?

Is there a change in context or location between these lines?

Is knowledge of Object A necessary for understanding the following line?

Does Line N have causal dependencies established in Line N-1?

Could Line N-1 occur before Line N?

A.4 Selected Questions

Does "awful" contradict an assertion on line 1?

Could "shaped" in line 4 be removed and the story world would remain unchanged?

Because of "tall" does line 9 contradict information in another line?

Could line 1 and 5 both be removed and have no effect on the story?

Is there a change in context or location between line 2 and 5?

Do the events in line 3 contradict events in line 2?