# Evaluating Analogy-Based Story Generation: An Empirical Study

**Jichen Zhu**
Digital Media
Drexel University
Philadelphia, PA USA
jichen@drexel.edu

**Santiago Ontañón**
Computer Science
Drexel University
Philadelphia, PA USA
santi@cs.drexel.edu

## Abstract

Evaluation is one of the major open problems in computational narrative. In this paper, we present an empirical study of *SAM*, an analogy-based story generation (ASG) algorithm, that was created as part of our *Riu* interactive narrative system. Specifically, our study focuses on *SAM*'s capability to retrieve and generate short non-interactive stories. Combining qualitative and quantitative methods from different disciplines, the methodology in this study can be extended to evaluating other computational narrative systems.

## Introduction

Computational narrative, especially story generation, is an important area for interactive digital entertainment and cultural production. Built on the age-old tradition of storytelling, algorithmically structured and generated stories can be used in a wide variety of domains such as computer games, training and education. This research area has made considerable progress in the past decades, notably in planning-based approaches (Meehan 1976; Lebowitz 1984; Riedl and Young 2004) and multi-agent simulation-based ones (Theune et al. 2003). New algorithmic improvements, often aided by narratology theories, have allowed computer systems to produce increasingly complex stories.

One of the major open problems for computational narrative is evaluation. When a system produces a story, how do we know how "good" the story is? An answer to this question would have at least two immediate ramifications. First, it would allow us to evaluate the system that produced these stories. In some cases, such as some forms of drama management (Weyhrauch 1997), the proper functioning of the system actually requires an automatic way to evaluate stories. Second, a set of well-designed evaluation methods will allow us to track progress in the field, and thus would be instrumental in articulating the overarching research directions in computational narrative.

However, it is tremendously difficult to evaluate computational narrative systems in terms of both the system performance and the narrative experience they provide. As Gervás observes, "[b]ecause the issue of what should be valued in a story is unclear, research implementations tend to sidestep it, generally omitting systematic evaluation in favor of the presentation of hand-picked star examples of system output as means of system validation" (Gervás 2009).

There has been increasing interest in evaluation in the interactive narrative community (Callaway and Lester 2002; Vermeulen et al. 2010; Schoenau-Fog 2011). Focusing on story generation systems, a subset of interactive narrative, we surveyed existing evaluation methods in these systems and categorized them into three main categories: 1) providing sample stories generated by the system, 2) evaluating the system processes, and 3) assessing the stories through user studies. In our past work, we further outlined several general evaluation principles for a more comprehensive evaluation methodology with mixed methods to better assess the user's narrative experience (Zhu 2012).

In this paper we present a specific case study of how we evaluate the analogy-based story generation algorithm SAM, part of our computational narrative system *Riu*. Informed by our prior work on the general issue of narrative evaluation and empirical literary studies, a highly related research area which has not been sufficiently explored in computational narrative, we design our evaluation approach in ways that further explore the users' reading experience. The work presented here extends a recent evaluation of SAM (Zhu and Ontañón in press) by further analyzing the user study data, specially in comparing user-provided analogical data with that generated by SAM. The results from this new study help us better understand the performance of our system as well as provide new insights into analogy-based story generation, pointing to promising future research directions.

## Related Work

In our survey of existing evaluation methods in story generation systems (Zhu 2012), we found most of them fall into into the following three broad approaches. First, providing sample stories generated by the system is one of the most common approaches for validating the system as well as the stories it generates. This approach started from the first story generation system *Tale-Spin* (Meehan 1981), where sample stories (translated from the logical facts generated by the system into natural language by the system author) are provided to demonstrate the system's capabilities as well as its limitations. Similarly, many later computational narrative systems such as *BRUTUS* (Bringsjord and Ferrucci 2000)

and *ASPERA* (Gervás 2000) also use selected system output for validation. Although this approach aligns with the tradition in literary and art practice to showcase the more interesting final work, "handpicking star examples" without stating the system author's criteria for selection can be potentially problematic.

The second approach is to evaluate the system based on its underlying algorithmic process. For instance, *Universe* (Lebowitz 1985) provides fragments of the system's reasoning trace, along with the corresponding story output. In this case, the story output is not the end goal. It signals that the *underlying process* of the system (i.e., learning) has a certain capacity (i.e., creativity) illustrated by the output.

The third approach, currently gaining more momentum, is *user studies*. For instance, *Minstrel* was evaluated through a series of user studies, regarding both the content and the presentation (e.g., better grammar and polished prose) of the generated stories (Turner 1993). Many of these studies are modeled after the Turing Test. In the *MEXICA* system (Pérez y Pérez and Sharples 2001), an internet survey is conducted about stories generated by different settings of *MEXICA* as well those generated by other systems. The users rated the stories along a 5-point Likert scale regarding five aspects of the stories: coherence, narrative structure, content, suspense, and overall experience. The authors of the *Fabulist* system (Riedl 2004) also used a similar approach.

The evaluation presented in this paper combines different aspects of all three approaches. We hand-selected SAM-generated stories that are representative along different similarity dimensions and story qualities. We then conducted a user study about readers' perception of the analogies between these stories, measured at each different stage of generative process. We evaluated different stages of our system's process with the goal of achieving better analogy-based story generation (ASG). However, we do not make any cognitive claims about our system. One of our main contributions is the incorporation of qualitative methods, such as grounded theory and the continuation test developed in the field of empirical literary studies.

## SAM and the *Riu* System

*Riu* is a text-based interactive narrative system designed to explore the connection between an external story world and the character's inner world of memories and imagination. In *Riu*, the user controls a character through a story world. The character might recall memories (which are displayed to the user) from the past upon encountering a situation analogous to one of those memories. Further, when the user enters a command to ask the character to perform an action, the character first *imagines* what will happen after performing such action, and if the imagined consequences are not appealing, might refuse to perform the action. Additionally, if the main character imagines an appealing outcome for one of the available actions, she might perform that action directly without user input. An in-depth description of *Riu* can be found in our prior work (Ontañón and Zhu 2010a).

This connection between the character's inner memory/imagination world and the external story world is modeled using computational analogy via two main functionalities: *story retrieval*, and *analogy-based story generation*. The former is used to allow the main character to retrieve memories similar to the situation at hand, and the latter to imagine the outcome of actions by analogy with some of the recalled memories. The study presented in this paper aims at evaluating these two functionalities.

### Structural Mapping & Story Representation

The Structure Mapping Engine (SME) algorithm (Falkenhainer, Forbus, and Gentner 1989) is the analogy-making algorithm used in *Riu*. The foundation of SME is Gentner's structure-mapping theory on the implicit biases and constraints by which humans interpret analogy and similarity (Gentner 1988). Built upon psychological evidences, Gentner argues that human analogical reasoning favors the relations between entities, rather than their surface features.

The story representation used in *Riu* is based on *force dynamics*. Force dynamics is a semantic category defined by cognitive linguist Leonard Talmy (Talmy 1988). It is based on the observation that a wide range of human linguistic and cognitive concepts are understood by considering them as if they were physical forces. A basic force dynamics (FD) pattern contains two entities: an *Agonist* (the focal entity) and an *Antagonist*, exerting force on each other. An Agonist has a *tendency* towards either motion/action or rest/inaction. It can only manifest its tendency if it is stronger than the opposing Antagonist. In *Riu*, unless specified otherwise, we represent the protagonist of the story as the Agonist. At the temporal level, Talmy uses a *phase* to describe the interaction between Agonist and Antagonist at a particular point in time. A story is represented as a sequence of phases, and each phase contains both a frame-based representation as well as natural language. In the remainder of this paper, we will use the term *scene* to refer to a small encapsulated piece of story, typically involving one main character in a single location. More detail of the FD framework can be found in (Ontañón and Zhu 2010b).

### Story Retrieval & Generation

Story retrieval refers to finding, amongst a set of predefined stories, one that is the most similar to a given target story. This is used in *Riu* for retrieving memories that are similar to the situation at hand. This is done via two steps:

1. *Surface Similarity*: Using a, computationally cheap, *keyword similarity*, select the $k$ stories that share the largest number of keywords with the target (in our study $k = 3$).
2. *Structural Similarity*: Then, SME is used to compute a, computationally expensive, structural similarity between the $k$ selected stories and the target. The story with the highest structural similarity is retrieved.

Story generation in *Riu* is performed via the SAM analogy-based story generation algorithm (Ontañón and Zhu 2011), and is used in *Riu* to "imagine" the consequences of a particular user-selected action in the story world by transferring knowledge from one of the recalled memories.

SAM takes two input parameters: $T$ and $S$ (the target and source scenes respectively), and outputs a new scene $R$, as the completion of $T$ by analogy with $S$. For the rest of the

paper we will say that an *analogical connection* is an individual one-to-one correspondence between a single entity or relation in the source domain and another one in the target domain, and use the term *mapping* as the complete set of connections found between the two domains.

SAM consists of four main steps (see (Ontañón and Zhu 2011) for a formal description):

1. **Temporal Mapping**: Generate the set $M$ of all possible phase mappings between the two input scenes.
2. **Analogical Mapping**: for each phase mapping $m \in M$, use SME to generate a mapping between all the story elements (characters, actions, etc.) of the target and those of the source, and to compute a *score*, which indicates how *good* the analogy is. The best phase mapping $m^* \in M$ and analogical mapping $g_{m^*}$ are selected.
3. **Resulting Scene Creation**: a new scene $R$ is created by appending to $T$ those phases from $S$ that were not part of the mapping $m^*$.
4. **Analogically Transform the Resulting Scene**: The reverse of the analogical mapping $g_{m^*}$ is applied to all the phases $R$, and then all the entities that came from $S$ and are not related to entities from $T$ are removed from $R$.

Thanks to the story representation used by SAM, the previous process returns both a frame-based representation of the story, as well as natural language. Below we show an example of SAM's output (this is S/T 2 in Table 1):

> [Source:] *Julian hadn't eaten all day and was counting on the crab trap to provide him with a feast of hearty shellfish. When he pulled the trap off the water it was filled with the largest crabs he had ever seen. So large in fact that the weight of them caused the rope to snap just before Julian could pull the trap onto the deck.*
>
> [Target:] *Zack is on deck, ready to face a storm. There's a flash of lightning in the distance. Suddenly, there's a bump on the side of the boat. Zack looks over. It is a gigantic cod! He's never seen one this large and close to the surface before. The storm is closing in. He races to get some fishing gear and try to catch it.*
>
> [SAM generated continuation:] *When Zack pulled the fishing gear off the water it was filled with the largest cod Zack had ever seen. So large in fact that the weight of cod caused the rope to snap just before Zack could pull the fishing gear onto the deck.*

To generate this story, elements from the source were mapped to elements in the target. For example, 'fishing gear' is mapped to 'crab trap.' In this case, SAM completed the story by adding one additional phase at the end.

## Evaluation: A User Study

To evaluate the effectiveness of the ASG components in our system, our user study focuses on answering the following three research questions: 1) How effective is the system in *identifying analogical connections* in ways similar to readers in our user group? 2) Is our choice of *force dynamics* for story representation suitable in the context of computational narrative? 3) What is the *quality of the stories* generated by our system from the perspective of the readers?

Table 1: Properties of four source-target (S/T) pairs.

| | S/T 1 | S/T 2 | S/T3 | S/T 4 |
|---|---|---|---|---|
| FD similarity | low | low | high | high |
| Surface similarity | low | high | low | high |

### Study Design

The scope of this study is to assess how well our system can generate short, non-interactive stories and whether these stories are aligned with a reader's intuitive notions of similarity and analogy. We position this study as the necessary foundation before we revise and evaluate *Riu* as a whole.

The study contains four main tasks, each of which evaluates one of our system's main generative steps. In each task, participants answer the same set of questions for four different source-target (S/T) story pairs, each containing a complete source story and an incomplete target story. These stories are excerpts from *Evening Tide*, an interactive story world created for *Riu*. The stories used in this study have a simple narrative structure: all the source stories contain two force dynamics phases, whereas all target stories contain one phase. The four S/T pairs represent four combinations of surface and structural similarity (i.e., force dynamics) levels (Table 1). The pairs are also selected because, from our perspective, the stories SAM generated using them represent a range of its performance. The average length of a source story is 73.25 words, and 38.0 for target stories. Finally, SAM-generated stories were minimally edited to fix capitalization and to add/remove missing/extra determinants.

In order to minimize our impact on the participants' interpretation of the stories, the user study instructions are kept minimal. The participants are only informed of the broad topic of the study—computational narrative; no information about the system or whether/which stories were generated are revealed. We also avoid any unnecessary technical jargon such as force dynamics. Source and target stories are simply referred to as Story A and B. In addition, we changed the name of the protagonist to a different one in each story in order not to imply any connection.

### Results

In response to our email recruitment, 31 people completed the survey. Among them, 27 are male, 3 female, and one undisclosed. Their age range was between 18 and 49, with a mean between 26 and 27. On average, each participant spent 35.20 minutes to complete the study. (We excluded one participant for computing this average since her data seems to indicate that she did not complete the survey in a single session.) Below are results for each task.

### Task 1 (Story Elements Mapping)

This task was designed to evaluate to what extent our system can identify mappings between source and target in ways similar to humans. For each S/T pair, a participant sees a source story, a target story and two lists of entities (i.e., characters and objects) and relations (e.g. "Herman is at the booth") explicitly mentioned in the source and the target sto-

Table 2: The fraction of participants who identified the same analogical connections as SAM under different configurations, and the average size of the mapping they found.

| | Random | Human | FD | bare | WN | FD+WN |
|---|---|---|---|---|---|---|
| S/T 1 | 0.04 | 0.46 | 0.48 | 0.35 | 0.35 | 0.48 |
| S/T 2 | 0.06 | 0.61 | 0.76 | 0.81 | 0.81 | 0.76 |
| S/T 3 | 0.05 | 0.57 | 0.48 | 0.47 | 0.48 | 0.48 |
| S/T 4 | 0.07 | 0.75 | 0.77 | 0.77 | 0.77 | 0.77 |
| Avg. | 0.05 | 0.60 | 0.63 | 0.60 | 0.61 | 0.63 |
| Size | - | 7.14 | 4.25 | 3.25 | 3.75 | 4.00 |

Table 3: Number of acceptable (Acc.) and erroneous (Err.) generated stories using the users found mappings and their respective average mapping size and connection score.

| | Acc. | Err. | Size of the mapping | Score |
|---|---|---|---|---|
| S/T 1 | 25 | 6 | 5.8 / 8.5 | 0.46 / 0.48 |
| S/T 2 | 14 | 17 | 7.5 / 9.59 | 0.67 / 0.56 |
| S/T 3 | 19 | 12 | 4.25 / 8.00 | 0.61 / 0.52 |
| S/T 4 | 8 | 23 | 7.13 / 7.61 | 0.80 / 0.72 |
| Avg. | 16.50 | 14.50 | 6.91 / 7.49 | 0.64 / 0.54 |

ries. Each participant is asked to identify as many analogical connections between the two lists as possible.

To assess the utility of force dynamics (FD), we compared the mappings identified by the participants with a set of randomly generated mappings, and with those identified by SAM with 4 different domain knowledge settings. The settings are a) *FD*: the default setting only with force dynamics; b) *bare*: we removed the FD annotations from the default setting; c) *WN (WordNet)*: we supplement the *bare* setting with domain knowledge of categories automatically extracted from the *WordNet*'s "hypern" database (for instance, the entity 'fish' is supplemented with *WordNet* properties such as 'aquatic-vertebrate,' 'vertebrate,' and 'animal.'); and d) *FD+WN*: the combination of the FD and WN settings.

A visualization of the connections found by our participants and by our system (with force dynamics) is shown in Figure 1, where, for each source-target pair we show two tables (one for entities and one for relations). Each row corresponds to an entity or relation in the target story, and each column to an entity or relation in the source story. Darker shades indicate that a higher percentage of participants identified such connection. Black squares mark the connections identified by our system. We can see that in S/T 2 and 4 (with a high FD structure similarity), both participants and our system tend to converge towards the same mapping. But the same is not true for S/T 3 and specially not for 1.

We computed the average fraction of participants that had identified each of the connections found by each configuration of our system, we call this the *connection score* (Results as shown in Table 2). Notice that the scores achieved by SAM are not very far from those by humans. The key difference between human participants and SAM is the size of the mappings (i.e., the number of analogical connections) they each find. The participants found an average of 7.14 connections, whereas SAM found fewer. FD, moreover, helps identifying more connections than *WordNet* does.

Table 4: Kendall $\tau$ correlation index between the ground truth and different configurations of *Riu*.

| | FD | bare | WN | FD+WN |
|---|---|---|---|---|
| Structural Similarity | 0.08 | 0.13 | 0.33 | 0.21 |
| Surface Similarity | 0.33 | 0.33 | 0.29 | 0.33 |
| Random Ordering | 0.50 | | | |
| Random Participant | 0.14 | | | |

Given that force-dynamics helps our system find more analogical connections, which in turn helps generating better stories, we performed an additional experiment. We used SAM to generate stories using the connections identified by each of the participants, instead of those found by SME. This experiment has the goal of evaluating whether finding more analogical connections, or connections with higher connection score results in better stories.

We classified the resulting stories by hand into those that were acceptable (*Acc.*), and those that contained big semantic mistakes (*Err.*). The classification was currently done by the authors, however, we plan to further validate it by more users. Then, we compared the connection score and the size of the mappings for each group. Results are shown in Table 3. Interestingly, SAM tends to generate erroneous stories when the size of the mapping is large. Analyzing the data, this occurred because when too many connections are identified, some are likely to map entities or relations that play very different roles in the stories, thus resulting in low quality output. Also, notice that the number of acceptable stories decreases from S/T 1 to S/T 4, since S/T 1 is the pair where the stories are less similar, and thus participants identified less connections in average.

The average connection score for those mappings that generated acceptable stories (0.64) is higher than for those that generated erroneous stories (0.54). It indicates that while it is important for an analogy-based story generation system to find large mappings between source and target, if those mappings are too large, the quality of the output might suffer. Thus, mappings should be limited to those entities and relations that play a very similar role in both stories.

## Task 2 (Story Similarity)

This task allows us to compare the stories that the participants find the most similar to a target story to our system's results. For each of the 4 S/T pairs, the participant is asked to rank 4 potential matching source stories based on their respective similarities to a target story. The participants' rankings of the potential matching stories were aggregated using the standard Borda count (Sandholm 1999). The aggregated participant's ordering, which we refer to as the *ground truth*, is compared with the ranking generated by *Riu*'s memory retrieval component. We do so by using the Kendall $\tau$ ranking correlation index (Kendall 1938), which is 0 for two identical orderings, 1 for opposite orderings, and expected to be around 0.5 for random orderings.

We compared the ground truth with: a) a random ordering, b) the ordering given by a random participant in our study, and c) the ordering *Riu* generated with only FD, with-

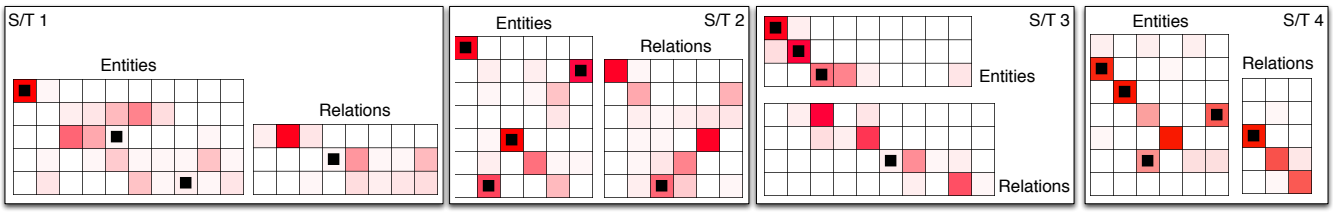


Figure 1: Shades indicate the percentage of users that identified each of the connections. Black squares mark those identified by our system using force dynamics.

out FD or *WordNet*, with only *WordNet*, and with both. In each domain knowledge setting, we tested *Riu* in two conditions: a) only using a basic surface similarity measure (based on the percentage of keywords shared between the two stories), and b) using both surface and structural similarity measures, as actually used in *Riu*. Results are summarized in Table 4. Again, the results show that using FD, our system obtained the best results (an ordering almost identical to the ground truth, with a $\tau$ distance of only 0.08). In addition, all the orderings generated using surface similarity are different from the ground truth, justifying *Riu*'s use of the computationally expensive structural similarity.

## Task 3 (Analogical Projection)

This task aims at evaluating the quality of SAM's analogical projection. For each S/T pair, the participant is presented with a complete source story and an incomplete target story. We first ask her to continue the incomplete story by writing at most 3 relatively simple sentences in English free-text. This method is based on what is known as the "story continuation test" from the Empirical Literary Studies field (Emmott, Sanford, and Morrow 2006).

Next, we present the participant with a continuation generated by SAM, along with the continuation she just wrote, and ask her to rate it on a 5-point Likert scale, Figure 2 summarizes these results, showing the means and standard deviations of the ratings. Among the 4 SAM-generated story continuations, S/T 2 and S/T 3 are considered relatively high quality by the participants (3.70 and 3.27), while the other two get lower ratings. The continuation rated the lowest was from S/T 1, quoted below:

> [Target:] *As a child, Eva would often sniff the honeysuckle in the backyard. Unbeknownst to her, there was a bee's nest by the honeysuckle.*
> [SAM-Generated Continuation:] *Eva cried.*

The problem is that, due to the low similarity between source and target, the analogical mapping generated by our system does not contain enough elements appearing in the second phase of the source, and thus, little could be transferred. As illustrated below, the participants faced the same difficulty in their own free-writing. S/T 4 with high surface and high FD similarity also received a lower rating, even though SAM transferred a lot of content from the source story, because SAM made a semantic mistake in the generated story.

The participants' own free-writing story continuations provided us with useful information to contextualize their

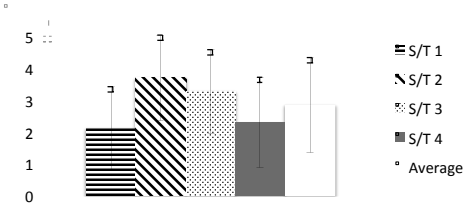


Figure 2: Average scores of SAM-generated continuations.

ratings of the ones SAM generated using the same S/T pairs. As used in other literary contexts, the continuation test gives a direct indicator regarding which aspects of the original stories a participant deems most important and hence continues them. This information offers a useful starting point for the design of future ASG systems. However, the open-ended nature of free-writing also presents a significant challenge for *qualitative* analysis. For each S/T pair, we clustered these continuations using the grounded theory method (Glaser 1992). Each continuation is coded based on which elements and relations are explicitly mentioned. We then iteratively cluster them until patterns start to form. Some of these patterns are related to our expected analogical connections. For example, one of the continuation to S/T 2 (above), "Zack let loose with his fishing rod. The storm rocked the boat and snapped his line before he could catch a fish," is categorized in a cluster for "tool broke & storm appeared." Other times, new categories emerge from the user's text. The continuation "The storm capsizes his ship," for instance, belongs to the "storm" cluster.

Our results are summarized in Table 5. The "Clusters" column shows the number of clusters obtained. In S/T pairs with low surface similarity (i.e., S/T 1 and 3), participants came up with a more varied set of continuations. By contrast, in S/T 4 (strong FD and surface similarity), most participants converge in how to continue the story, resulting in very few clusters. All participants' continuations are also divided into two groups based on whether they are analogous to the source or not. For story pairs with high FD similarity (S/T 3 and 4), a significant number of participant-authored continuations are analogous to the source. In particular, for S/T 4, 24 out of all 29 participant-authored continuations share a strong analogy with the source. Moreover, in story pairs with low FD similarity (S/T 1 and 2) participants wrote continuations that were free formed (and thus, not analogous

Table 5: Analysis of the participants' free-writing.

| | Clusters | Analogous to Source | Not Analogous |
|---|---|---|---|
| S/T 1 | 6 | 19 | 11 |
| S/T 2 | 5 | 16 | 14 |
| S/T 3 | 7 | 23 | 6 |
| S/T 4 | 3 | 24 | 5 |

to the source) much more often. This is evidence that participants used the source story to generate the continuation more often when they found a clear analogy (in our case, when the two stories had a similar FD structure). This hints at a very interesting direction for future research, for example analogy can be used as the main story generation mechanism, but it can be complemented (when no good analogy can be found) with other generation methods, such as planning (like (Riedl and León 2009)). Another idea is to retrieve additional sources, when a good analogy cannot be found with the current one.

An interesting observation is that our participants were not necessarily consistent between the different tasks. For example, one participant found a very good analogical mapping for S/T 2 (with 10 connections), but then produced a continuation ("The storm capsizes his ship") that did not leverage most analogical connections he identified in the previous task. Further, we observe that some of the more interesting story continuations are those which follow analogical rules and then inserted new narrative elements within some reasonable boundaries. This finding, combined with the results shown in Table 3, indicate that the analogical mapping found should only be used as a *guidance* for story generation. The algorithm should be flexible enough to modify the mapping based on new elements introduced. This idea has been explored by Gunes et al. (2012).

### Task 4 (Overall Story)

Finally, this task evaluates the quality of the complete stories generated by SAM. In addition to the four story continuations generated by SAM (also used in Task 3), we added a low-quality ASG story, created by manually copy-pasting the second phase of a story after the first phase of another, and one completely human-authored story. These two additions are intended to set a baseline for the range of scores. The 6 stories are rearranged randomly for each participant. Each story is rated by participants on a 5-point Likert scale along three dimensions: plot coherency, character believability, and overall quality.

Results, summarized in Figure 3, show that the ratings for the low-quality story and the human-authored story define the two extremes and set the context for the rest. The scores obtained by SAM are closer to those of the human-written story than to the low-quality one. Specially in terms of character believability, SAM's score was relatively high (3.88 on average out of 5 compared to 4.43). Certain generated stories, 2 and 3, obtain much higher scores than the average. Comparing Figure 2 and Figure 3 we can see that the overall scores obtained by the stories in Task 4 are highly correlated with those in Task 3, as expected. The difference between

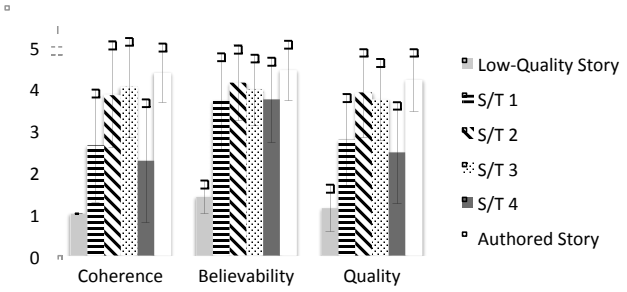


Figure 3: Average scores of SAM-generated Stories in Task 4, compared to 2 benchmark stories.

SAM's scores and the low-quality story were found to be statistically significant using a paired t-test with $p < 0.05$. The difference between SAM's scores and the human-authored story were found statistically significant with $p < 0.05$, except for stories 3 and 2 for coherence and quality.

## Conclusions

In this paper we present an empirical study of the analogy-based story generation algorithm SAM in the *Riu* System. The evaluation has been designed to answer three research questions, to which we obtained the following answers: 1) our system finds analogical mappings that align with those found by our participants; as a consequence, the retrieval mechanism also aligns with the participants' notion of similarity. 2) Force dynamics clearly helps improving the performance of our system. 3) Finally, the quality of stories generated by SAM is still not on par with a human authored story, but some o them were rated relatively highly.

The evaluation of our system combines three main evaluation approaches in computational narrative: we select representative sample stories generated by SAM, then evaluate the underlying algorithmic processes through these samples, and finally, we do so via a user study. Additionally, we designed our user study as an attempt to incorporate much-needed qualitative methods, such as grounded theory and the continuation test from the literary studies community. Further integration of quantitate and qualitative methods based on their respective strengths in different aspects of computational narrative evaluation is part of our future work.

We believe that the basic principles behind our study design can be generalized to evaluate other story generation systems. By selecting a small set of representative stories generated by the system, and using them, in the context of a user study, to analyze each of the different algorithmic steps the system performs, we provide a step towards a more balanced evaluation method. Additionally, variations of our continuation test can be used to compare the performance of the system with that of human participants under similar constraints. As part of our future work, we plan to scale up our evaluation to include the complete interactive *Riu* system, and generalize the principles behind our evaluation methodology, to be applicable to a wide range of story generation systems.

## References


Bringsjord, S., and Ferrucci, D. A. 2000. *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. Hillsdale, NJ: Lawrence Erlbaum.

Callaway, C. B., and Lester, J. C. 2002. Narrative prose generation. *Artificial Intelligence* 139(2):213–252.

Emmott, C.; Sanford, A.; and Morrow, L. 2006. Capturing the attention of readers? stylistic and psychological perspectives on the use and effect of text fragmentation in narratives. *Journal of Literary Semantics* 35(1):1–30.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41:1–63.

Gentner, D. 1988. Structure-mapping: A theoretical framework for analogy. In *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*. San Mateo, CA: Kaufmann. 303–310.

Gervás, P. 2000. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems* 14:200–1.

Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49–62.

Glaser, B. G. 1992. *Emergence vs forcing: Basics of grounded theory analysis*. Sociology Press.

Gunes Baydin, A.; Lopez de Mantaras, R.; and Ontañón, S. 2012. Automated generation of cross-domain analogies via evolutionary computation. In *Proceedings of the Third International Conference on Computational Creativity (ICCC-12)*, 25–32.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30(1/2):81–93.

Lebowitz, M. 1984. Creating characters in a story-telling universe. *Poetics* 13:171–194.

Lebowitz, M. 1985. Story-telling as planning and learning. *Poetics* 14(6):483–502.

Meehan, J. 1976. *The Metanovel: Writing Stories by Computer*. Ph.d., Yale University.

Meehan, J. 1981. Tale-spin. In *Inside Computer Understanding: Five Programs Plus Miniatures*. New Haven, CT: Lawrence Erlbaum Associates.

Ontañón, S., and Zhu, J. 2010a. Story and Text Generation through Computational Analogy in the Riu System. In *AIIDE 2010*, 51–56. The AAAI Press.

Ontañón, S., and Zhu, J. 2010b. Story representation in analogy-based story generation in riu. In *Proceedings of IEEE-CIG 2010*, 435–442.

Ontañón, S., and Zhu, J. 2011. The SAM Algorithm for Analogy-Based Story Generation. In *AIIDE*, 67–72. The AAAI Press.

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(2):119–139.

Riedl, M., and León, C. 2009. Generating story analogues. In *AIIDE 2009*. The AAAI Press.

Riedl, M. O., and Young, R. M. 2004. An intent-driven planner for multi-agent story generation. In *Proceedings of AAMAS 2004*, 186–193. IEEE Computer Society.

Riedl, M. 2004. *Narrative Generation: Balancing Plot and Character*. Ph.D. Dissertation, North Carolina State University.

Sandholm, T. W. 1999. Distributed rational decision making. In Weiss, G., ed., *Multiagent systems*. Cambridge, MA, USA: MIT Press. 201–258.

Schoenau-Fog, H. 2011. Hooked! evaluating engagement as continuation desire in interactive narratives. In *Proceedings of the Fourth International Conference on Interactive Digital Storytelling (ICIDS 2011)*, 219–230.

Talmy, L. 1988. Force dynamics in language and cognition. *Cognitive Science* 12(1):49–100.

Theune, M.; Faas, E.; Nijholt, A.; and Heylen, D. 2003. The virtual storyteller: Story creation by intelligent agents. In *Proceedings of TIDSE 2003*, 204–215.

Turner, S. R. 1993. *Minstrel: a computer model of creativity and storytelling*. Ph.D. Dissertation, University of California at Los Angeles, Los Angeles, CA, USA.

Vermeulen, I. E.; Roth, C.; Vorderer, P.; and Klimmt, C. 2010. Measuring user responses to interactive stories: towards a standardized assessment tool. In *Interactive Storytelling*. Springer. 38–43.

Weyhrauch, P. 1997. *Guiding Interactive Drama*. Ph.D. Dissertation, Carnagie Mellon University.

Zhu, J., and Ontañón, S. in press. Shall I compare thee to another story: An empirical study of analogy-based story generation. *IEEE Transactions on Computational Intelligence and AI in Games*.

Zhu, J. 2012. Towards a new evaluation approach in computational narrative systems. In *Proceedings of the Third International Conference on Computational Creativity (ICCC 2012)*, 150–154.