Designing An Interdisciplinary User Evaluation for the *Riu* Computational Narrative System

Jichen Zhu

Drexel University, Philadelphia, PA, USA
jichen.zhu@drexel.edu

Abstract. Evaluation is one of the major open problems in Interactive Digital Storytelling (IDS) research. As narrative systems grow in their capacities, the community needs a set of well-designed evaluation methods and criteria that can bring insights on the systems as well as the stories they provide. In this short paper, we examine existing evaluation methods in the area of generative narrative system, and identify several important properties of stories and reading that have so far been overlooked in empirical studies. We present our preliminary work of developing a more interdisciplinary evaluation approach that takes into account both the system and cultural aspects of the computational narrative system *Riu*.

1 Introduction

Evaluation is one of the major open problems in IDS research. A set of well-designed evaluation methods not only is instrumental in informing the development of better computational narrative systems, but also helps to articulate overarching research directions for the field over all. 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 in the context of computational narrative, "[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 [2]."

This is not an isolated phenomenon, but occurs across many computational research areas that intersect with cultural and creative domains such as music and the visual arts. A recent survey of 75 creative systems shows that, only slightly above half of the related publications give details on evaluation; among those, the main aim and evaluation criteria are quite different [4].

We argue that the difficulty of establishing evaluation methodology in computational narrative, research reflects the cultural clash between the scientific and the arts/humanities practices. Aligned with Snow's notion of the two cultures, many researchers active in the intersection of both communities have observed their different and sometimes opposing value systems and axiomatic assumptions [6, 16, 19]. For example, Simon Penny [12] argues, in the context of digital media

art, that sciences insistence upon alphanumeric abstraction, logical rationality, and desire for generalizability are fundamentally at odds with the affective power of artwork, which is based on specificity and complexity. In the context of evaluation, this conflict takes the form of the clash between the productivity-and-value-based methodology adopted by both AI and HCI communities, and the general resistance to empirical studies in the arts and humanities.

In this short paper, we present our preliminary work of developing a more interdisciplinary evaluation approach that takes into account both the system and cultural aspects of computational narrative systems. We built on our initial work [18] and present our user study design. Our work is not intended to replace the function of literary criticism and close reading with simple empirical studies and statistical analysis. We also believe that evaluation is a critical process to inform the development of narrative systems and to deepen the understanding of how to provide new forms of narrative experiences. In the rest of paper, we first examine existing evaluation methods in computational narrative, focusing on story generation systems, and identify several important properties of stories and reading that have so far been overlooked in existing evaluations. Drawn upon methods from empirical literary studies, we then present our preliminary work on designing user evaluation studies on our computational narrative system, *Riu*.

2 Existing Framework on Narrative Evaluation

This section provides an overview of existing evaluation methods in story generation systems. Some of our observations can also be applied to interactive digital storytelling systems in general. Recent examples of evaluating the latter type can be found in [17, 15]. Based on our survey of major text-based story generation systems, existing evaluation methods can be grouped into three categories.

2.1 System Output Samples

As Gervás pointed out above, providing sample generated stories 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* [7], 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. In addition to successful examples, Meehan also picked different types of "failure" stories to illustrate the algorithmic limitation of the system for future improvement. Similarly, many later computational narrative systems such as *BRUTUS* [1] use selected system output for validation. One reason for the wide use of this approach is its alignment with traditional literary and art practice, where the final artifact should stand on its own without formal evaluation beforehand. However, simply showing the "successful" output without stating the system author's criteria for selection can be potentially problematic. Some more recent work in this approach has attempted to make this selection process more transparent. For example, using

the WordNet knowledge base, the authors of the *Riu* system developed a measure of semantic distance to evaluate the quality of the analogy generated by their system [10].

2.2 Evaluating System Process

The second approach is to evaluate the system primarily based on its underlying algorithmic process. For instance, the *Universe* system [5] provides fragments of the system's reasoning trace, along with the corresponding story output, in order to show how the underlying process leads to the particular output. This category often contains systems that use narrative to illustrate/model underlying cognitive processes. For example, the author shows an example of how *Universe* learn new "plot fragments" by generalizing from given example stories. The bigger research goal is to illustrate the system's capability to expand its plot-fragments library automatically, and hence the learning process is a necessary condition to creativity.

2.3 User Studies

Minstrel is evaluated by a series of user studies in order to determine the quality of the stories it generates. In the first user study, 10 users were asked to read the generated stories, without being told that they were generated by a computer, and to answer questions regarding their impression of the author and the story. In the second study, 10 users repeated the above test, except the generated stories were rewritten by a human writer for better presentation with improved grammar and more polished sentences. In the third study, the same questions were asked about another story written by a 12-year-old as a benchmark.

A larger number of users were involved in the evaluation of the *MEXICA* system [13]. An Internet survey about the generated stories was sent out and 50 users submitted their answers. The users rated 7 stories by answering a set of 5-point Likert scale questions over five factors (i.e., coherence, narrative structure, content, suspense, and overall experience). Among these 7 stories, 4 were generated by *MEXICA* using 4 different system configurations (with or without certain modules). Two stories were generated by other computational narrative systems (*GESTER* and *MINSTREL*). In the *Fabulist* system [14], the system author conducted two quantitative evaluations. The first one is to evaluate plot coherence: a story is shown to different users; each of them then rate the importance of each sentence in the story, based on the assumption that unimportant sentences decrease plot coherence. Second, character believability in the stories is evaluated by asking users to rate the difference in characters' motivation in stories generated by two configurations of the system.

3 User Study Design

There are three aspects, among other things, that we need to address in a more culturally driven user study of computational narrative systems. First, the user

study needs to acknowledge different audiences and different modes of reading. For example, an ordinary user will be more likely to adopt story-driven reading, which focuses more on the immersiveness of the stories. They contemplate what characters are doing, experience the stylistic qualities of the writing, and reflect on the feelings that the story has evoked [9]. An expert reader, on the other hand, will more readily adopt the point-driven orientation. They perform informed close reading — a complex act of interpretation at the linguistic, semantic, structural, and cultural levels — in order to understand the "point" of plot, setting, dialogue, etc. These qualitative expert-novice differences have long been acknowledged in the literary empirical studies of linear text, and should be incorporated into evaluations of computational narrative.

Second, evaluations of the narrative experience provided by computational systems need to be measured against system and content authors' intention. In many of the evaluations we surveyed above, system output is evaluated based on either a set of cross-system criteria, such as character believability and plot logical coherence, or on how much readers enjoy the stories. Although these criteria provide useful milestones for the research community, it is important not to forget the assumptions built in these criteria, that is, they embody the quality of the narrative that the system authors *intend* to create. Storytelling is, after all, a form of communication between the author and the reader. In some cases, the authors may intend to focus more on the emotional atmosphere created by the system, rather than plot coherence. In other cases, a user's report of unpleasantness may be positive or even desirable, if the system author intends to use her stories to challenge the reader's belief system, in ways similar to Duchamp's Urinal. In other words, evaluation criteria of specific narrative systems should take into account the particular expressive goals of their authors.

Third, as a whole field, we will benefit from more mixed approach that use both quantitative and qualitative methods. A large percentage of the evaluations we surveyed gravitate towards quantitative methods with qualitative methods as a supplement, if at all. Through surveys and experiments, numerical data is collected, then analyzed statistically to provide an average user response. Although these methods have the clear advantage of being relatively easy to collect and analyze, they filter out the specificity and contextualization that is crucial to cultural artifacts. More details of the discussion can be found in [18].

3.1 Study Design Guidelines

The computational narrative system we plan to evaluate is the *Riu* system [11]. It uses computational analogy to generate a text-based interactive narrative experience about a character's internal activities such as memories and daydreams. Through analogical retrieval and analogical projection, these internal activities are used to enrich and influence the "physical" world of the character. For instance, while encountering an object in the "physical" world, the character may retrieve memories of similar objects, which will in return change his disposition towards it and hence possible actions.

We are primarily interested in the narrative effect of adopting the parallel structure between the character's "physical" world and inner world brought forth by computational analogy. In other words, as our first step, we intend to understand whether and to what extent the internal activities of the characters affect an ordinary reader's (hence story-driven reading orientation) emotional connection with the main character. As system authors, our intention is to create a new kind of interactive narrative experience that focuses on association (i.e., similarities between objects and events) rather than cause-and-effect (as in many planning-based computational narrative systems). It is more important, to us, if our system creates memorable narrative moments and evokes deep emotions than providing logical and coherent plots. As a result, we will not evaluate our system based on "plot coherence" or "character believability." Instead, our study will center around readers' connection with the character and their general emotional response to the stories. For the kind of rich exchange of meanings that *Riu* intends to evoke, quantitative data captured by Likert scale questionnaire alone is not sufficient to capture the rich interpretive process people engage in reading. Our study seeks to supplement quantitative data with qualitative open-ended interviews. As a result, the study is geared less towards statistical significance of the users we include, but rather the depth of the response of each user. Overall, the users will be randomly assigned into two groups. One of them will interact with the system with the analogy-driven internal activities and the other group without. Their interaction with the system will be video recorded and the participant will be interviewed with retrospective protocol for their experience. This general methodology has been used in Façade [8] and art-oriented digital systems [3].

More specifically, in order to gain insights into how memorable the interactive narrative experience is to each user, we will adopt a recall test. Each user, after completing their interaction with the system, will be asked to perform comprehension and recall tasks. For instance, the user will be asked to recall as many phrases and story elements they read as possible. Although both tasks are well-developed methods in understanding reader response, often used in empirical literary studies, to the best of our knowledge they have not been substantially used in the evaluation of computational narrative systems. By asking users to answer specific questions and recall phrases from the story, we hope to gather more reliable data about how engaged the users are in the story than simply asking them to rate the experience. It will also allow us to compare the effect of incorporating character's internal activities between the two groups.

4 Conclusion

In this short paper, we discussed the challenge of designing evaluation methods for interactive narrative systems. Based on our survey of existing approaches, we identified three main aspects we hope to address in order to better understand interactive stories as expressive cultural artifacts. Drawing upon methods in empirical literary studies, we presented our preliminary design for the user

evaluation of our analogy-based computational narrative system that is geared towards the above three main aspects.

References

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

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

[3] Höök, K., Sengers, P., Andersson, G.: Sense and sensibility: evaluation and interactive art. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 241–248 (2003)

[4] Jordanous, A.: Evaluating evaluation: Assessing progress in computational creativity research. In: Proceedings of the Second International Conference on Computational Creativity (ICCC-11). pp. 102–107 (2011)

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

[6] Mateas, M.: Expressive ai: A hybrid art and science practice. Leonardo 34(2), 147–153 (2001)

[7] Meehan, J.: Tale-spin. In: Inside Computer Understanding: Five Programs Plus Miniatures. Lawrence Erlbaum Associates, New Haven, CT (1981)

[8] Mehta, M., Dow, S., Mateas, M., MacIntyre, B.: Evaluating a conversation-centered interactive drama. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems. pp. 8:1–8:8 (2007)

[9] Miall, D.S., Kuiken, D.: Foregrounding, defamiliarization, and affect response to literary stories. Poetics 22, 389–407 (1994)

[10] Ontañón, S., Zhu, J.: On the role of domain knowledge in analogy-based story generation. In: Proceedings of the Twenty-Second International Joint Conferences on Artificial Intelligence (IJCAI-2011). pp. 1717–1722 (2011)

[11] Ontan, S., Zhu, J.: Story and text generation through computational analogy in the riu system. In: Proceedings of AI and Interactive Digital Entertainment Conference (AIIDE 2010). pp. 51–56. AAAI Press (2010)

[12] Penny, S.: Experience and abstraction: the arts and the logic of machines. In: Proceedings of PerthDAC 2007: 7th Digital Arts and Culture Conference (2007)

[13] Pérez y Pérez, R., Sharples, M.: Mexica: A computer model of a cognitive account of creative writing. Journal of Experimental & Theoretical Artificial Intelligence 13(2), 119–139 (2001)

[14] Riedl, M.: Narrative Generation: Balancing Plot and Character. Ph.D. thesis, North Carolina State University (2004)

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

[16] Sengers, P.: Anti-Boxology: Agent Design in Cultural Context. Ph.D. thesis, Carnegie Mellon Universit (1998)

[17] Thue, D., Bulitko, V., Spetch, M., Romanuik, T.: A computational model of perceived agency in video games. In: Proceedings of the Seventh Conference on Artificial Intelligence and Interactive Digital Entertainment. pp. 91–96 (2011)

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

[19] Zhu, J., Harrell, D.F.: Navigating the Two Cultures: a Critical Approach to AI-based Literary Practice, pp. 222–246. World Scientific, Singapore (2011)