Towards a Mixed Evaluation Approach for Computational Narrative Systems

Jichen Zhu
Drexel University
Philadelphia, PA 19104 USA
jichen.zhu@drexel.edu

Abstract

Evaluation is one of the major open problems in computational creativity research. Existing evaluation methods, either focusing on system performance or on user interaction, do not fully capture the important aspects of these systems as cultural artifacts. In this position paper, we examine existing evaluation methods in the area of computational narrative, and identify several important properties of stories and reading that have so far been overlooked in empirical studies. Our preliminary work recognizes empirical literary studies as a valuable resource to develop a more balanced evaluation approach for computational narrative systems.

Introduction

Evaluation is one of the major open problems in computational creativity research. A set of well-designed evaluation methods not only is instrumental in informing the development of better creative computational systems, but also helps to articulate overarching research directions for the field overall. However, research in creative systems has encountered tremendous difficulties in defining suitable evaluation methods and metrics, both at the level of individual systems and across systems. A recent survey of 75 creative systems shows that, only slightly above half of the related publications give details on evaluation; among those, there is lack of consensus on both the aim of evaluation and the suitable evaluation criteria (Jordanous 2011).

Traditionally, methods for evaluating intelligent computational systems have been mainly developed in two areas: artificial intelligence (AI) and human-computer interaction (HCI). Following the scientific/engineering tradition, evaluation in AI typically relies on quantitative methods to measure the system's performance against a certain benchmark (e.g., system performance, algorithmic complexity, and the expressivity of knowledge representation). A salient example is the measure of "classification accuracy" in machine learning, where new algorithms are evaluated by being compared to standard ones over the same sets of data. Whereas the AI community is primarily concerned with the operation of the system itself, HCI concentrates on the interaction between the user and the system. Borrowing from psychology, human factors, and other related fields, HCI developed a set of quantitative and qualitative *user study* methods to understand the usability of a system along such principles as learnability, flexibility, and robustness (Dix et al. 2003).

Although these existing approaches offer useful insights into creative systems as functional and useful products, they do not fully capture a crucial property of creative systems, that is, they are and they produce cultural artifacts such as stories, music, and paintings. In these areas, there has not be an established tradition of formal evaluation. When we combine artistic expression and system building, evaluation becomes an issue. 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" (2009).

We argue that the difficulty of establishing an evaluation methodology in computational creativity research reflects the cultural clash between the scientific/engineering and the humanities/arts practices. Aligned with Snow's notion of the two cultures (1964), researchers working in the intersection of the two communities have observed the conflict of different and sometimes opposing value systems and axiomatic assumptions (Mateas 2001; Sengers 1998; Manovich 2001; Harrell 2006; Zhu and Harrell 2011). One of the differences is what Simon Penny (2007) calls the "ontological status of the artifact" between the electronic media arts practice and computer science research. For an artwork, the effectiveness of the immediate sensorial effect of the artifact is the primary criterion for success. As a result, most if not all effort is focused on the persuasiveness of the experience, which is built on specificity and complexity. In computer science, the situation is reversed. The artifact functions as a "proof of concept" and hence its presentation can be overlooked; the real work is inherently abstract and theoretical. These differences, Penny argues, illustrate that the insistence upon "alphanumeric abstraction," logical rationality, and desire for generalizability in science is fundamentally at odds with the affective power of artwork. In the context of evaluation, this conflict takes the form of the clash between the productivity- and value-based methodologies adopted by both AI and HCI communities, and the general resistance to empirical studies in the arts.

In this position paper, we present our initial work of developing a more balanced evaluation approach that takes into

account *both system and cultural* aspects of creative systems, focusing on computational narrative systems and their output. Our work is not intended to replace the function of literary criticism and close reading with empirical studies and statistical analysis. Simplistic attempts to reproduce art as a scientific experiment without an in-depth understanding of the former's tradition and value systems are shortsighted (as discussed in Ian Horswill's panel presentation at the Fourth Workshop on Intelligent Narrative Technologies, Palo Alto, 2011) and counter-productive to the long-term goal for computational creativity research. In the meantime, we also believe that evaluation is a critical process to inform the development of creative systems and to deepen the understanding of computational creativity. Therefore, more research and discussions about evaluation are needed.

In the rest of paper, we examine existing evaluation methods in the area of computational narrative, and identify several important properties of stories and reading that have so far been overlooked in existing evaluation methods. Our preliminary work suggests that empirical literary studies can be a valuable resource to develop a more balanced evaluation approach for computational narrative systems.

Existing Work on Narrative Evaluation

Broadly speaking, discussions of evaluating creative systems have taken place at two levels. At the level of computational creativity in general, researchers have attempted to come up with domain-independent evaluation criteria to measure a system's level of creativity, both in terms of its process and output. For example, Colton (2008) and Jordanous (2011) proposed standardized frameworks to empirically evaluate system creativity. The importance of these approaches is that, in addition to evaluating specific systems, they also allow potential cross-domain comparison between systems. At the level of specific creative domains, evaluations are conducted to validate a specific creative system and its output in that domain. For instance, the recent work by Vermeulen et al. in the IRIS project (2011) proposed a list of standardized, systematic assessment criteria for interactive storytelling systems using concepts that "play a key role in users' responses to interactive storytelling systems."

This section provides an overview of existing evaluation methods in the area of computational narrative. Our main focus is on the evaluation of story generation systems and their output, but some of our observations can also be applied to (non-generative) interactive digital storytelling systems. Recent examples of evaluating the latter type can be found in (Thue et al. 2011; Schoenau-Fog 2011). Although we do not specifically deal with high-level constructs such as 'novelty' and 'value,' we believe that more comprehensive evaluation criteria at the domain-specific level can indirectly contribute to the recognition and formulation of these high-level creativity constructs at the first level. Based on our survey of major text-based story generation systems, existing evaluation methods can be categorized into three broad approaches.

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* (Meehan 1981), where sample stories (translated from the logical propositions generated by the system into natural language by the system author) are provided to demonstrate the system's capabilities. 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* (Bringsjord and Ferrucci 2000), and *ASPERA* (Gervás 2000) use selected system output for validation. Besides the lack of established specific evaluation metrics, the reason for the wide appeal of this approach is that it aligns with the tradition in literary and art practice where the final artifact should stand on its own without formal evaluation.

However, simply showing the "successful" and/or "interesting" output without explicitly stating the system author's selection criteria can be potentially problematic. Some recent work in this approach has attempted to make this selection process more transparent. For example, in the evaluation of the *GRIOT* system, Harrell (2006) evaluates the generated poems based on the quality and novelty of the metaphors they invoke. When the system generates "my world was so small and heavy," the author evaluates it by the metaphor it evokes — "Life is a Burden." Similarly, the *Riu* system (Ontañón and Zhu 2011) automatically assesses the generated stories by measuring the semantic distances of the analogies in the stories based on the WordNet knowledge base.

Evaluating the System's Process

The second approach is to evaluate the system primarily based on its underlying algorithmic process. Among the three evaluation approaches, this one is most aligned with traditional AI evaluation methods. Cognitive systems often use this approach to show that the system's underlying processes are cognitively sound. For instance, the evaluation of the *Universe* system (Lebowitz 1985) included fragments of the system's reasoning trace, along with the corresponding story output. It is intended to illustrate the system's capability to expand its plot-fragments library by generalizing from given example stories. Although the sample output and the process are relatively simple compared to those of the previous approach, Lebowitz intends to show, especially through the system processes, that the learning process is a necessary condition to creativity.

In a more complex example, the *Minstrel* system (Turner 1993), presented as a model for the creative process and storytelling, is evaluated in two ways. First, Turner evaluates the system by comparing it to related work in psychology, creativity, and storytelling. *Minstrel*'s process is contrasted to existing AI models of creativity both in the similar domain of narrative (e.g., *Tale-Spin* and *Universe*) and in different ones (e.g., AM (Lenat 1976)). Second, *Minstrel* is empirically studied in terms of its plausibility and quality as a test

bed for evaluating different hypotheses of creativity. Specifically, plausibility is evaluated based on 1) the quantity of possible output stories, by testing the system in different domains, and 2) the quality of output stories through a series of user studies (details in the next section). In the evaluation of the "test bed" criteria, Turner studies why some TRAMS (i.e., problem-solving strategies) were added, removed, etc. to prove that one can experiment with different models of creativity. For instance, to test its model of "boredom" as how many repeated elements are there in the stories, *Minstrel* was asked to generate stories about the same topic four times. The differences and similarities between these stories are analyzed to evaluate how boring these stories are.

User Studies

Evaluating the system's process alone, however, does not provide insights into the quality of the output. For systems that are more geared towards seeing narrative as a goal in its own right, user studies provide a way to assess the output story without counting solely on the author's own intuition. As a result, user studies has been increasingly adopted both as a standalone evaluation method and as a complement to other approaches.

For example, the *MEXICA* system (Pérez y Pérez and Sharples 2001) is evaluated through an Internet survey. The users rated seven 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 seven stories, four were generated by *MEXICA* using different system configurations (with or without certain modules). Two stories were generated by other computational narrative systems (i.e., *GESTER* and *MINSTREL*). The last story was written by a human author using "computer-story language." The scores each stories received is used to determine *MEXICA*'s level of "computerised creativity" (c-creativity) in reference to human writers and other similar systems.

In a more complex example, in addition to the methods mentioned above, the stories generated by *Minstrel* are evaluated through a series of independent user studies. In the first user study, users were given the generated stories, without being told that they were generated by a computer. Then they were asked to answer questions regarding their impressions of the author and the stories. In the second study, a different group of users repeated the above test, except the generated stories were rewritten by a human writer for better presentation with improved grammar and more polished prose. In the third study, the users were presented an unrelated story written by a 12-year-old and asked to answer the same set of questions.

User studies of narrative systems do no always adopt some form of the Turing Test. In the *Fabulist* system (Riedl 2004), the system author conducted two quantitative evaluations without using human writers as a benchmark. The first study evaluates plot coherence, measured based on the assumption that unimportant sentences decrease plot coherence. A group of users independently rates the importance of each sentence in the generated story and hence the coherence of the plot. Second, character believability is evaluated by asking users to rate the difference in characters' motivation in stories generated by two configurations of the system.

What is Missing

Computational narrative is still in its early stage, both in terms of the depth and breath of the narrative content. It is especially true when we compare these generated stories with what we typically conceive as literary text produced by human authors. In this regard, the different methods described in the previous section are arguably adequate for the current state of these systems. As argued above, however, evaluation methods play an important role not only in assessing existing systems, but also in informing what kind of future systems should be built. In this regard, waiting for the narrative systems to mature before starting to develop suitable evaluation criteria is detrimental to the research community.

As computational narrative research moves forward, a set of more comprehensive evaluation methods can help to reduce the gap between computer generated stories and traditional literature. *Our position is that* many important lessons from literary criticism and communication theory are by and large overlooked in computational narrative. We argue that they can be instrumental to developing evaluation methods that not only focus on the algorithmic and usability aspects of narrative systems, but also the expressiveness of the generated stories as cultural artifacts.

Below is our preliminary work in identifying some crucial elements that are missing in many existing evaluation methods. It is not intended to be seen as a comprehensive list, but rather as an initial step towards incorporating *fundamental* knowledge and concerns from related fields in the arts and the humanities.

Different Modes of Reading

Reading is a complex activity. Depending on the setting, purpose of the reading, and background of the reader, different aspects of the text are highlighted. Vipond and Hunter (1984) distinguished among point-driven, story-driven, and information-driven orientations for reading. Shown by recent studies in Reader Response theory (Miall and Kuiken 1994), ordinary readers typically adopt the story-driven approach, that is, to read for plot. They contemplate what characters are doing, experience the stylistic qualities of the writing, and reflect on the feelings that the story has evoked. This mode is adopted while we read for pleasure.

By contrast, the point-driven orientation is the foundation for literary criticism. Experts 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. Point-driven reading assumes that the text is a purposeful act of communication between the author and the reader, and the "points" in the story have to be constructed through the reader's careful examination of the text.

Finally, in the information-driven orientation, a reader is more concerned about extracting specific knowledge from the text. We adopt this orientation while, for example,

following a recipe or checking facts in an encyclopedia. Information-driven reading places a strong emphasis on the coherence and informativeness of the text. This orientation is less common in computational narrative.

Different reading orientations place different emphasis on evaluation methods. As story-driven reading is primarily concerned with creating the "lived-through experience" for the reader, compatible evaluation needs to focus on the immersiveness of the story world. In computational narrative, most existing evaluation criteria presume the story-driven reading orientation and center on interestingness, presence, and engagement of the stories (e.g. plot coherence and character believability). Additionally, this orientation requires the participants of the evaluation to be close to an "average reader." A point-driven-based evaluation requires participants, usually experts, to perform more in-depth reading of the text beyond the surface plot. The effectiveness of different literary techniques, such as thematic structures, linguistic patterns, and points of view in the story can be evaluated in ways similar to traditional literary criticism.

To the best of our knowledge, there have not been attempts of point-driven-based evaluation in the context of computational narratives. There are many complex reasons for this. Some may argue that computational narrative, at its current stage, is too simple for this level of close reading. However, electronic literature (e-lit) work demonstrated that less algorithmically complex systems can still produce rich meanings. Establishing these evaluation criteria helps to develop a wider range of computational narrative.

Authorial Intention

Contradictory to the tradition of literary criticism, the evaluation of computational narrative systems has by and large ignored the intention of the authors. If we subscribe to the assumption that storytelling is a form of communication between the author and the reader, authorial intention should play a role in evaluating how effective these stories are. For instance, 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. A more balanced evaluation needs to differentiate this scenario from unpleasantness caused either by poorly written story or by unintuitive user interface. Similarly, intentional ambiguity in the story can be a powerful device, leaving something undetermined in order to open up multiple possible meanings. In the history of literature, intentionally ambiguous works such as Henry James's 1898 novel *The Turn of the Screw* have triggered many distinctive interpretations and vigorous debates about them.

Mixed 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.

Several research projects have attempted to address this issue. Mehta et al. (2007) devised an empirical study for the *Façade* system, which was intended by its authors to evoke rich exchange of meanings. Mehta et al. acknowledge that the standard quantitative criteria in the conversational system research community (e.g., task success rate, turn correction ratio, concept accuracy and elapsed time) are not adequate because they assume a task-based philosophy, where conversational interaction is framed as a simple exchange of clear, well-defined meanings. As a result, they made a deliberate choice to use more in-depth but less statistically significant ethnographic methods to study a small group of users' perceptions and interpretations of their conversations with non-player characters. Using video recording and retrospective interviews, their study found that participants created elaborate back-stories to make sense of character reactions in order to fill in the gaps of AI failures, an insight difficult to capture with pure quantitative methods.

The limitation of quantitative methods is echoed in Höök, Sengers and Andersson's user study of their digital art project (Höök, Sengers, and Andersson 2003). They observed, "[g]rossly speaking, the major conflict between artistic and HCI perspectives on user interaction is that art is inherently subjective, while HCI evaluation, with a science and engineering inheritance, has traditionally strived to be objective. While HCI evaluation is often approached as an impersonal and rigorous test of the effects of a device, artists tend to think of their system as a medium through which they can express their ideas to the user and provoke them to think and behave in new ways." As a response, their interpretive methods (open-ended interviews) focuses on giving the artists a grounded feeling for how the interactive system was interpreted and their message was communicated. Despite the sentiment against user studies in the interactive arts community, some artists involved in the project acknowledged that laboratory evaluations can help artists to uncover problems in interaction design.

Because of these limitations, we believe that a mixed methods approaches may be more suitable for evaluating computational narrative outputs. In addition to the closed-ended questions and surveys, qualitative methods such as phenomenology, grounded theory, ethnography, case studies can better capture the plurality of meanings interpreted by different readers and the complexity of such readings.

In literary studies, a group of researchers have started developing methods to empirically study readers' responses to literature. Due to the field's predisposition to point-driven interpretation, these methods offer a good example of balancing expert interpretation and ordinary readers' responses to and experience of the stories under evaluation. For example, Miall (2006) identified four kinds of empirical literary studies. First, studies that manipulate a literary text to isolate a particular effect. Second, studies that use an intact text in which the researchers hypothesize that intrinsic features of the text influence the reader. Instead of manipulating a text, each text itself provided a naturally varying level of foregrounding from high to low. A third kind of study involves comparison of two or more texts. Four, readers are asked to think aloud about a text during or after reading it. All of

these can be further explored and potentially incorporated into the evaluation of computational narrative systems.

Conclusion

In this position paper, we discussed the challenge of designing evaluation methods for creative systems due to their dual status. Focusing in the area of computational narrative, we surveyed existing evaluation approaches in story generation systems and identified crucial aspects of computational narrative, as a potential form of cultural artifacts, that have been so far downplayed. Penny warned us of the danger of the "unquestioned axiomatic acceptance of the concept of generality as being a virtue in computational practice especially when that axiomatic assumption is unquestioningly applied in realms where it may not be relevant" (Penny 2007). We suggest that work in empirical literary study research can offer valuable insights of developing more interdisciplinary and more balanced evaluation methods.

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.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI 2008 Spring Symposium in Creative Intelligent Systems*. AAAI Press.

Dix, A.; Finlay, J.; Abowd, G.; and Beale, R. 2003. *Human-Computer Interaction*. Edinburgh Gate, England: Prentice Hall.

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.

Harrell, D. F. 2006. Walking blues changes undersea: Imaginative narrative in interactive poetry generation with the griot system. In Liu, H., and Mihalcea, R., eds., *AAAI 2006 Workshop in Computational Aesthetics: Artificial Intelligence Approaches to Happiness and Beauty*, 61–69. Boston, MA: AAAI Press.

Höök, K.; Sengers, P.; and Andersson, G. 2003. Sense and sensibility: evaluation and interactive art. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 241–248.

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

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

Lenat, D. B. 1976. *Am: an artificial intelligence approach to discovery in mathematics as heuristic search*. Ph.d., Stanford University.

Manovich, L. 2001. Post-media aesthetics, available at http://www.manovich.net/docs/post_media_aesthetics1.doc.

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

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

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

Miall, D. S., and Kuiken, D. 1994. Foregrounding, defamiliarization, and affect response to literary stories. *Poetics* 22:389–407.

Miall, D. S. 2006. *Literary Reading: Empirical and Theoretical Studies*. NewYork: Peter Lang.

Ontañón, S., and Zhu, J. 2011. 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)*, 1717–1722.

Penny, S. 2007. Experience and abstraction: the arts and the logic of machines. In *Proceedings of PerthDAC 2007: 7th Digital Arts and Culture Conference*.

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

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

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.

Sengers, P. 1998. *Anti-Boxology: Agent Design in Cultural Context*. Ph.D. Dissertation, Carnegie Mellon Universit.

Snow, C. P. 1964. *The Two Cultures*. New York: Menton Books.

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

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.; Roth, C.; Vorderer, P.; and Klimmt, C. 2011. Measuring user responses to interactive stories: Towards a standardized assessment tool. In *Proceedings of the Fourth International Conference on Interactive Digital Storytelling (ICIDS 2011)*, 38–43.

Vipond, D., and Hunt, R. A. 1984. Point-driven understanding: Pragmatic and cognitive dimensions of literary reading. *Poetics* 13:261–277.

Zhu, J., and Harrell, D. F. 2011. *Navigating the Two Cultures: a Critical Approach to AI-based Literary Practice*. Singapore: World Scientific. 222–246.