STORYEVAL: An Empirical Evaluation Framework for Narrative Generation

Jonathan P. Rowe[1] **Scott. W. McQuiggan**[2] **Jennifer L. Robison**[1]

Derrick R. Marcey[1] **James C. Lester**[1]

[1] Department of Computer Science, North Carolina State University, Raleigh, NC, USA
[2] Education Practice, SAS Institute, Inc., Cary, NC, USA
{jprowe, jlrobiso, drmarcey, lester}@ncsu.edu, scott.mcquiggan@sas.com

Abstract

Research in intelligent narrative technologies has recently experienced a significant resurgence. As the field matures, devising principled evaluation methodologies will become increasingly important to ensure continued progress. Because of the complexities of narrative phenomena, as well as the inherent subjectivity of narrative experiences, effectively evaluating intelligent narrative technologies poses significant challenges. In this paper, we present STORYEVAL, an evaluation framework for empirically studying computational models of narrative generation. Drawing on evaluation methodologies from cognitive science, human-computer interaction, and natural language processing, as well as techniques that have begun to emerge in the narrative technologies community, STORYEVAL consists of four complementary tools for evaluating both interactive and non-interactive narrative generation: Narrative Metrics, Cognitive-Affective Studies, Director-centric Studies, and Extrinsic Narrative Evaluations. We discuss the benefits and limitations of each family of techniques and illustrate their application with example narrative generators drawn from the field.

Introduction

Recent years have seen significant growth in research on intelligent narrative technologies. Much of this work has focused on narrative generation (Turner 1994; Riedl and Young 2005), and work on non-interactive narrative generators has sought to capture an array of complex narrative phenomena, ranging from character intentionality (Riedl and Young 2005) to models of the creative process (Turner 1994). Interactive narrative systems have used similar generative techniques to balance well-formed story experiences with significant player agency. Their capacity to dynamically construct and revise story plans in response to users' actions has shown promise for applications in education (Zoll et al. 2006; Mott and Lester 2006), training (Si, Marsella, Pynadath 2005), entertainment (Cavazza, Charles, and Mead 2002; Riedl, Saretto, and Young 2003; Magerko 2007), and art (Mateas and Stern 2005). These systems have leveraged a variety of computational approaches to narrative generation and drama management, including adversarial search (Weyhrauch 1997), planning (Cavazza, Charles, and Mead 2002; Riedl, Saretto, and Young 2003), decision-theoretic approaches (Si, Marsella, Pynadath 2005; Mott and Lester 2006), and Markov decision processes (Nelson et al. 2006; Roberts et al. 2006).

Interactive narrative systems have won accolades for novel story generation technologies (Mateas and Stern 2005), they have been embraced by international audiences of hundreds and thousands of users (Johnson and Beal 2005; Mateas and Stern 2005; Zoll et al. 2006), and they have been used as effective educational tools across many domains (Johnson and Beal 2005; McQuiggan et al. 2008). Despite these successes, progress in the field has proven difficult to measure for several reasons. First, narrative experiences are intrinsically subjective, making critical assessment notoriously unreliable. Second, narratives are typically authored for specific domains and content, which makes it difficult to systematically compare alternate computational models and the systems that embody them. Third, narratives are enormously complex, multi-dimensional constructs. Despite thousands of years of storytelling, refinement, and analysis, there does not (and perhaps cannot) exist any canonical theory of narrative. For scientists and engineers seeking to advance the field's state-of-the-art, these factors pose significant challenges.

In this paper, we address these issues by proposing an empirical evaluation framework for studying computational models of narrative generation. Our hope is that by highlighting the most effective approaches for assessing narrative generation, we can better explore the role of empirical evaluation in intelligent narrative technologies. Further, by analyzing the techniques used to evaluate narrative generators, we can obtain a clearer view of the community's progress, the methods necessary for measuring it, and insights into the most promising approaches for accelerating advancements in the field.

We begin with a brief overview of narrative and the challenges inherent in evaluating narrative generators. We then present STORYEVAL, an empirical narrative generation evaluation framework that draws on methodologies from cognitive science, human-computer interaction, and natural language processing, as well as the narrative technologies community itself. We describe each of STORYEVAL's four components (Narrative Metrics, Cognitive-Affective Studies, Director-centric Studies, and Extrinsic Narrative Evaluations), discuss benefits and limitations, and suggest how they might be applied to specific narrative generators.

Characteristics of Intelligent Narrative Technologies

Intelligent narrative technologies encompass a broad range of techniques for story understanding, generation, interaction, and authoring. In this paper, we focus primarily on narrative generators. Following an overview of narrative and the philosophical perspectives that have informed work to date on narrative generation, we discuss the challenges of narrative technology evaluation.

Narrative Generation

Although narrative is generally defined as the representation of one or more events, such a simple definition fails to indicate the complexity that characterizes narrative phenomena. For example, narrative events often have intricate temporal and causal relationships, maintain one or more continuant subjects, and constitute a well-formed whole (Prince 2003). Critical analysis must consider issues of dramatic tension, plot structure, and character. Narratologists distinguish three components of narrative: fabula, sjuzet, and medium. The *fabula* is the story, consisting of the "set of narrated situations and events in their chronological sequence" (Prince 2003). The *sjuzet* is the discourse, the "set of narrated situations and events in the order of their presentation to the receiver" (Prince 2003). The *medium* is the delivery mechanism used to present the discourse to an audience, such as text, oral storytelling, animation, or film. Although an author may not consciously consider narrative in these terms as they craft a story, every complete narrative implements the components in some form.

Such reductionist approaches have had a significant influence on work in intelligent narrative technologies. Different projects have focused on different components of narrative, each accounting for specific sets of goals, requirements, and constraints. Further, as researchers have begun to build systems that incorporate interactivity into narrative, system goals and priorities have experienced a corresponding shift (Riedl, Saretto, and Young 2003). Introducing interactivity affects issues of fabula, sjuzet, and media. It places additional demands on character, and it carries important implications for constructing dramatic and well-structured stories. Consequently, both interactive narratives and non-interactive narratives pose their own distinct evaluation challenges.

Further complicating matters, intelligent narrative technologies have adopted several different philosophical approaches to constructing narrative. For example, researchers have distinguished between story-centric, author-centric, and world-centric approaches to narrative generation (Bailey 1999). *Story-centric* approaches view narrative as an abstract artifact with intrinsic characteristics to guide generation, perhaps best exemplified by story grammars. *Author-centric* narrative generation attempts to explicitly model human authors' story creation processes. *World-centric* approaches populate story worlds with autonomous characters, and then allow narrative to emerge from character interactions (Bailey 1999). Of course, each approach has its own strengths and weaknesses, and its own standards for assessment.

This complex landscape of alternative models and modalities calls for an amalgamation of complementary techniques for effectively assessing intelligent narrative technologies. Unfortunately, traditional human-computer interaction approaches are often ill suited to evaluating narrative phenomena. In the next section, we discuss why this is the case, and introduce techniques that are effective for narrative evaluation.

Evaluation Challenges

Evaluating narrative generators differs substantially from evaluating traditional software and AI systems. Classical AI models (e.g., theorem provers, planners) often use objective measures such as soundness, completeness, and optimality to assess a model's performance on a given task (Russell and Norvig 2003). Although narrative generators have a specific task, namely, to construct a story, the subjectivity and complexity inherent in narrative, as well as the sheer space of possible narratives, renders analysis of optimality and completeness difficult, if not impossible. Evaluation methodologies must consider which components (fabula, sjuzet, or media) the generator is targeting, how dependent and sensitive the resulting narrative is on the hand-authored specifications provided to the generator, what measures of "goodness" are appropriate for the stories generated, and for what aim or purpose were the generated stories created. Computational properties such as time and space complexity, robustness, and the space of possible stories are also important for assessing a generator, but these issues are usually overshadowed by concerns about the internal structure and surface presentation of the generated narratives.

Assessing interactive narrative generators also differs from more traditional software evaluation (e.g., database systems, word processors, e-mail clients). Several of the prominent assessment techniques used by the human-computer interaction (HCI) community employ analyses such as cognitive walkthroughs, heuristic evaluations, and model-based techniques (Dix et al. 2004). These approaches must make certain assumptions about the software being evaluated: the software enables the

completion of some well-defined task(s); its goals include ease-of-use, efficiency, and learnability; it behaves deterministically in response to user input; and existing cognitive models accurately reflect mental processing during user interaction. While these assumptions are appropriate for a wide range of systems, many of them break down when applied to narrative technologies. Interactive narratives often exhibit mixed-initiative, stochastic behavior; they may seek to intentionally prolong or frustrate a user for narrative effect; and emotion often influences user behavior as much as cognitive factors, which are not accounted for by GOMS and keystroke-level models. Further, factors such as character believability, plot coherence, dramatic tension, and narrative structure, all irrelevant to traditional software systems, are central to any heuristic analysis of a generated narrative. For these reasons, purely analytical approaches are often of limited value for assessing intelligent narrative technologies.

Another major approach to evaluation used by the HCI community is the user participant study. Currently, empirical approaches offer more promise for assessing narrative generators than purely analytical techniques. Unfortunately, human participant studies can be expensive. They also raise many practical issues regarding choice of participants, experimental design, logistics of laboratory and field studies, and statistical analysis of results. Nevertheless, empirical evaluation addresses many of the shortcomings associated with analytical approaches, so it is a widely used approach for assessing intelligent narrative technologies.

The issues that distinguish the evaluation of intelligent narrative technologies from other types of evaluation are reminiscent of those encountered by other AI sub-disciplines. Natural language processing is one example. Language is inextricably tied to narrative. It shares narrative's complex and multi-faceted nature, and its assessment is often subjective. These properties exacerbate the problems of evaluation. Work on embodied conversational agents (ECAs) has also raised challenging evaluation issues. The complexity of evaluating ECAs' natural language and dialogue behavior, as well as their capacity for expressive multimodal communication, complicates assessment. Fortunately, both fields have made significant progress in developing principled evaluation methodologies (Walker et al. 1997; Cassell et al. 2000; Belz and Reiter 2006), a cause for optimism for narrative generation evaluation.

An Empirical Evaluation Framework

Because narrative generators are complex systems, multiple methodologies must be employed to successfully evaluate the full scope of their functionalities and the stories and interactive experiences they create. To this end, we propose STORYEVAL, an empirical evaluation framework for computational models of narrative generation. The STORYEVAL framework consists of four complementary tools for empirically assessing interactive and non-interactive narrative generators:

- *Narrative Metrics*: By measuring specific characteristics of a generated narrative, narrative metrics can be used to evaluate the product of a narrative generator.
- *Cognitive-Affective Studies*: By gauging audience response to a narrative experience, cognitive-affective studies assess the impact of a narrative generator through human participant experiments.
- *Director-centric Studies*: By evaluating the computational performance of a director agent or drama manager, director-centric studies assess the effectiveness of a narrative generator.
- *Extrinsic Narrative Evaluations*: By assessing the performance of the application in which a narrative generator is embedded, extrinsic narrative evaluations measure the degree to which a narrative generator contributes to the application's overall effectiveness.

STORYEVAL integrates these four broad approaches into a single framework for comprehensively assessing narrative generators. The proposed methodology is neither automated nor algorithmic, but it provides a set of assessment techniques that can be adapted to individual interactive and non-interactive systems. We discuss each of STORYEVAL's four families of evaluation in turn.

Narrative Metrics

Narrative metrics focus assessment on the results produced by narrative generators. Unfortunately, there are no accepted, objective metrics for evaluating narrative artifacts; if there were, film and literary critics would be out of jobs. Instead, narrative metrics can leverage simple heuristics or user participant studies for assessment. Because of the lack of objective measures, and because there are too many variables associated with a comparison of machine-generated stories and human-generated stories, it can be difficult to measure machine-generated narrative against human standards. Instead, experimental designs can compare machine-generated stories to other machine-generated stories and determine the effects of various architectural components on narrative generation. Factors that have been assessed using this type of analysis include character believability (Riedl and Young, 2005) and narrative prose quality (Callaway and Lester 2001).

Narrative metrics can leverage empirically grounded theories from the social sciences. For example, Riedl and Young (2005) use a novel experimental approach that takes advantage of a well-grounded psychological model of story understanding, QUEST (Graesser et al. 1991). The system being evaluated, Fabulist, uses a variant of partial-order causal link planning to produce narratives that account for character intentionality. Riedl and Young conducted a human participant study that compared two versions of the Fabulist system: one uses an advanced planner (IPOCL), and the other uses a traditional partial order causal link planner. The two conditions compared participant judgments on generated narrative question-answer pairs against assessments provided by the QUEST models. The

experiment assessed participants' comprehension of character intentionality and Fabulist's ability to motivate character actions through the IPOCL-enhanced narrative generator. The investigators concluded that the enhanced narrative generator more effectively supports reader comprehension of character intentionality, although the novel and complex evaluation approach introduced some experimental design issues into the assessment.

Metrics such as style, readability, grammar, and diction can be used to evaluate narratives expressed in natural language. For example, AUTHOR is a narrative generator that combines story generation facilities and deep natural language generation to construct high quality narrative prose (Callaway and Lester 2001). AUTHOR is composed of five principal components: a discourse history, sentence planner, reviser, lexical choice component, and surface realizer. A human participant experiment was conducted to assess the system's generative performance. The system was provided with two different story plans, and was run on each with various architectural components removed. Conditions included no reviser, no lexical choice component, no discourse history, all three components working, and all three components disabled. This resulted in ten generated narratives, which were read and quantitatively graded by a pool of readers on narrative metrics such as style, readability, grammar, and diction. The study led the authors to conclude that the discourse history and revision components were particularly important to resulting narrative quality, with results concerning lexical choice being less conclusive. While this type of study could not compare machine-generated narrative against human standards, it was able to determine which sub-processes of narrative generation were important for producing quality results.

In addition to narrative metrics' use in a post-hoc manner in evaluation, they can also be incorporated directly into narrative generators. For example, the drama managers presented in Weyhrauch (1997) and Nelson *et al*. (2006) have used narrative metrics in the form of objective evaluation functions to assess candidate narrative directions. The functions were designed to declaratively encode authors' aesthetic preferences, against which narratives will be judged. The evaluation functions combine several measurements that are hypothesized to reflect authorial goals, such as spatial locality of action, topical locality of action, and the degree to which plot points are motivated by prior events. Combining common authorial goals into a single comprehensive, weighted measure, an objective evaluation function is then used during the optimization process that guides narrative decision making. This approach is useful for making rapid, simple assessments about the quality of a narrative or narrative experience, and is particularly attractive for generators that use machine learning or other optimization-based approaches. Unfortunately, simple evaluation functions are limited in their ability to measure many of narrative's most fundamental components. Further, the generality of the assumptions that associate particular narrative features with actual narrative "goodness" may be questionable. While more sophisticated, automated techniques could be implemented, the associated computational costs may violate real-time performance requirements.

Cognitive-Affective Studies

The quality of a narrative is inseparably tied to an audience's response to it. Cognitive-affective studies shift the focus away from narrative artifacts and toward the cognitive-affective states fostered by a narrative experience. While some experiments such as those discussed above can assess audiences' cognitive responses to a generated narrative, their focus is primarily on narrative-dependent metrics such as prose quality and character believability rather than on participants' emotional and attentional states.

Because one of the most powerful effects of narrative is the sense of being transported into a story (Gerrig 1993; Green and Brock 2000), narrative is well suited to fostering high levels of audience engagement and presence (Kelso, Weyhrauch, and Bates 1993; Rowe, McQuiggan, and Lester 2007). The fundamental premise motivating cognitive-affective studies of narrative is that if stories produced by narrative generators elicit responses that are reminiscent of those resulting from human-generated stories, the narrative generator is capable of producing quality narratives.

Unfortunately, it can be difficult to observe and assess cognitive-affective state. Many experiments request periodic emotion self-reports throughout a narrative experience (Lee 2007) or administer validated questionnaires following the completion of the intervention (McQuiggan, Rowe, and Lester 2008). Unfortunately, both of these techniques are highly subjective. Self-reports can jarringly interrupt a narrative experience, and post surveys take measurements long after cognitive-affective responses actually occur. Alternative techniques include facial expression analysis (Ekman 2003) and monitoring physiological measures such as heart rate and galvanic skin response (Lee 2007). However, most physiological measures only provide indirect indicators of cognitive-affective state. Despite their limitations, user participant studies hold much appeal for narrative generation evaluation.

Research on interactive narrative generators has long been interested in *presence*, informally defined as a user's sense of "being there" when interacting with a mediated environment (Schubert, Friedmann and Regenbrecht 1999; Insko 2003). Experiments investigating interactive narrative generators have yielded a number of surprising and interesting presence-related results. In some of the Oz group's earliest work, Kelso *et al*. (1993) investigated the notion of dramatic presence by observing a user participating in an interactive drama populated with live actors. They concluded that by being an active participant in the narrative, rather than a passive observer, the interactor "found interactive drama more powerful, easily

causing immediate, personal emotions, not the traditional vicarious empathy for other characters" (Kelso, Weyhrauch, Bates 1993). These experiments informed the work pursued by the Oz group over subsequent years. Unfortunately, the expense associated with using live actors make these types of experiments difficult to reproduce or to run on multiple participants.

McQuiggan *et al*. conducted a pair of experiments investigating the relationship between character behavior and user presence in an implemented interactive narrative (2008). The experiments compared two versions of CRYSTAL ISLAND, an interactive, 3D science mystery in which students learn about microbiology as they simultaneously discover the source of a mysterious illness plaguing the island. Both versions featured the same narrative, characters, world, and content, but one included a small subset of the characters who engaged users in short empathetic exchanges. Using a validated instrument for measuring presence, Witmer and Singer's PQ (1998), the studies found an increase in presence among students in the empathetic character condition. This result was produced across two populations, middle school and high school students, and it suggested that simple variations in character behavior can yield significant gains in user presence.

However, the relationship between presence and engagement in interactive narrative generators is not entirely clear, as evidenced by work from Dow *et al*. on an augmented reality version of Façade (2007). A human participant experiment compared an augmented reality version of Façade (AR Façade) with traditional desktop versions of the popular interactive drama. It was found that AR Façade elicited higher levels of presence than desktop versions. However, qualitative interviews conducted after the intervention found that the enhanced presence experienced in AR Façade did not correspond to increased levels of engagement. Some participants actually preferred the desktop version of Façade and indicated that they would rather "portray a character on the screen, rather than literally be in the situation" (Dow et al. 2007). The investigators hypothesized that the augmented reality interface made users feel "too close" to the socially uncomfortable scenario that Façade implements. This work suggests that while presence and engagement are important variables for assessing narrative experiences, narrative's objective is not necessarily a simple optimization of the two factors.

Integrally related to presence and engagement are assessments of emotional experiences fostered by narrative events. Many of the most powerful narrative experiences are defined by the affective responses they invoke: the horror genre seeks to elicit fear, comedies elicit joy, and the action genre elicits excitement. In recognition of the centrality of affective response in narrative, numerous human participant studies have been conducted to model and assess emotional responses to narrative interventions. For example, experiments with CRYSTAL ISLAND have combined emotional self-report data with physiological measures of heart rate and galvanic skin response to accurately model and assess emotional states during a narrative interaction (Lee, McQuiggan, Lester 2007). Other work on CRYSTAL ISLAND has focused on the transitions between different emotional states. Experimental evidence suggests that different types of empathetic character behaviors during a narrative interaction result in different emotion transition responses (McQuiggan, Robison and Lester 2008).

Director-centric Studies

Evaluation that centers on director agents and drama managers constitutes the third technique for evaluating narrative generation. Director agents themselves must perform significant narrative evaluation in the course of generating narratives. Director agents seek to provide well-formed narrative experiences, and in interactive narratives provide significant player agency (Riedl and Young 2003). To accomplish this objective, director agents should ideally consider the full scope of narrative—these include story elements such as plot, discourse, media, character, and drama—as well as expected user cognitive-affective responses, and then use the results to guide narrative decision making. Currently, most director agents perform a subset of these analyses, leveraging automated narrative metrics and cognitive-affective models to determine appropriate courses of action and intervention. It should be noted that narrative generation tasks need not be performed by a single centralized agent, but can be realized in a distributed manner, as is done in character-centric narrative generation (Cavazza, Charles, and Mead 2002). In this case, individual agents perform an additional form of metacognitive processing of their own actions (e.g., assessing emotional impact on others), and use this information to further guide behavior (Aylett and Louchart 2008). Regardless of approach, it is incumbent upon director agents to effectively and automatically evaluate current and potential narrative directions, and then use this information to manage the interactive narrative experience.

Director-centric studies can be used to assess the efficacy of particular strategies for balancing player agency and narrative structure, such as proactive intervention (Magerko 2007), reactive intervention (Riedl, Saretto, and Young 2003), and computational models of narrative rationality (Mott and Lester 2006). Some of the earliest evaluation work that centered on drama manager performance was conducted to assess the Moe architecture, which investigated three variants of adversarial search as mechanisms for informing a drama manager's narrative decision making (Weyhrauch 1997). Moe was run against nine different classes of simulated users, each varying in skill and cooperative tendency. For each model, user simulations compared search-enhanced interactive drama experiences against a version lacking a drama manager. It was found that the search-enhanced manager's resulting narrative distribution was significantly superior to the version lacking the manager, as measured by an aesthetic

evaluation function. However, Weyhrauch noted important limitations of his assessment: the evaluation focuses on the performance of the drama manager's search algorithm, rather than on the interactive drama as a whole, and it does not provide findings that can inform design improvements for a single experience. Moreover, the study did not include judgments provided by human users.

Although the director-centric evaluation technique for evaluating the Moe architecture did not include judgments solicited from human participants, the technique of comparing different narrative director implementations using simulated users is a promising one for preliminary assessment. Work at Georgia Tech by Nelson *et al*. (2006) and Roberts *et al*. (2006) has continued this line of research by comparing alternative optimization-based approaches for drama management. These projects have modeled the task of finding effective drama management strategies as reinforcement learning and Targeted Trajectory Distribution-MDP problems, respectively. Emphasizing the goal of affording significant user agency, their work highlights the importance of optimizing for a distribution of different, high quality stories, rather than merely focusing on policies that direct users toward a small set of highly rated narrative experiences. Determining the most effective strategies for achieving this goal remains an open research question.

Real-time performance constraints are another important consideration when evaluating narrative director agents. Often, the inherent narrative decision-making processes presented to a director agent are intractable. Weyhrauch addressed this problem by limiting Moe's search depth, as well through memoization strategies during online search (1997). Techniques used at Georgia Tech have simply moved the optimization process off-line (Nelson et al. 2006; Roberts et al. 2006). Mott and Lester's U-Director system implements a decision-theoretic approach to narrative management, a technique that poses a compute-intensive Bayesian inference problem during each narrative decision-making cycle (2006). To address this issue, they empirically investigated a number of different approximation techniques for Bayesian inference, the techniques' associated performance within the domain, and the effectiveness of their resulting decisions for guiding users through the narrative. Although U-Director's empirical evaluation was limited in scope, the findings underscored the importance of evaluating computational efficiency and its tradeoffs for narrative effectiveness.

Extrinsic Narrative Evaluation

The first three families of evaluation methodologies operate with an "inward facing" focus: they do not consider narratives' larger motivating contexts. To round out the evaluation framework's assessment methodologies, the final technique, extrinsic narrative evaluation, operates with an "outward facing" focus. Most narratives do not merely aim to recount a sequence of events; rather, they are used to entertain, communicate an idea, or serve some external purpose. For example, MINSTREL (Turner 1994) generates stories that communicate a theme or moral lesson. Façade (Mateas and Stern 2005) seeks to deliver an artistically complete, conversation-driven, dramatic experience. A number of interactive narratives generators aim to balance user agency and narrative coherence solely for the purpose of entertainment (Cavazza, Charles and Mead 2002; Riedl, Saretto and Young 2003). CRYSTAL ISLAND (McQuiggan, Rowe, and Lester 2008), FearNot! (Zoll et al. 2006), and the Tactical Language and Culture Training System (Johnson and Beal 2005) use narrative to contextualize learning and problem-solving scenarios. Extrinsic narrative evaluation is needed to assess narrative generation with an eye toward a narrative's purpose. The distinction between the "inward facing" and "outward facing" techniques play roles analogous to intrinsic and extrinsic evaluation in natural language processing (Jurafsky and Martin 2008). Intrinsic evaluations measure models independently of any particular application, while extrinsic evaluations assess models within an application and gauge the application's overall effectiveness. We discuss several evaluation approaches that measure narrative generators by their ability to produce narratives that support some extrinsic goal.

Extrinsic evaluation is critical for narrative generators used in the service of education and training. Educational narratives naturally lend themselves to extrinsic evaluation. These applications provide measureable variables that can be used to assess the overall performance of a narrative system, such as learning gains. Recently, intelligent narrative technology research teams have begun to collaborate with colleagues in the learning sciences to conduct user participant studies. For example, laboratory studies investigating the CRYSTAL ISLAND narrative-centered learning environment have shown significant learning gains among eighth graders after a single interaction with the science mystery (McQuiggan et al. 2008). Field studies involving the FearNot! narrative learning environment, which targets social education about bullying, have investigated changes in students' empathetic characteristics after completing the narrative scenario (Zoll et al. 2006). Researchers building the Tactical Language and Culture Training System have completed a number of iterative usability and learning evaluations in conjunction with the US Army (Johnson and Beal 2005).

A narrative generator need not serve an educational purpose to benefit from extrinsic evaluation. For example, Mehta *et al.* (2007) performed a qualitative evaluation of Façade's conversational system within the context of its larger dramatic objective. The authors ran several human participants through Façade and focused their attention on points where the system's conversational facilities failed. The authors concluded that Façade was relatively successful at maintaining user engagement and sense of drama. Curiously, users would often interpret conversational breakdowns as natural features of the narrative, and they inferred that conversational cues and character responses generated by Facade were an important part of these experiences.

Major drawbacks associated with extrinsic evaluation include the expense of embedding narrative technologies into full applications and the difficulty of conducting large, controlled human participant studies with appropriate populations. Clearly, extrinsic evaluation requires the existence of reasonably mature systems. Nevertheless, when extrinsic evaluation is possible, it can be an effective means for assessing narrative technologies.

Discussion and Conclusions

The STORYEVAL framework represents a first step toward an integrated evaluation methodology for computational models of narrative generation. By employing narrative metrics, cognitive-affective studies, director-centric studies, and extrinsic narrative evaluations, we can systematically assess precisely which aspects of a narrative generator most effectively contribute to its successful performance. STORYEVAL offers a promising beginning for a comprehensive narrative generation evaluation framework, but it does not address the evaluation of related narrative tasks such as story understanding or narrative authoring. Nevertheless, it highlights a number of central issues for evaluating intelligent narrative technologies:

- The complexity inherent in intelligent narrative technologies calls for a sophisticated multi-faceted approach to evaluation.
- While narrative generation evaluation methodologies can draw on techniques from cognitive science, human-computer interaction, and natural language processing, the assessment of narrative generation raises issues that are fundamentally different from those found in other types of software design.
- Narrative metrics, cognitive-affective studies, director-centric studies, and extrinsic narrative evaluations are integrally interrelated, and each has its own benefits and limitations.
- Narrative evaluation is not merely important for empirical validation, but its techniques can also form the basis for computational models of narrative generation.

As evidenced by progress in natural language processing, adopting effective evaluation methodologies can facilitate the rapid advancement of a field (Belz and Reiter 2006; Walker et al. 1997), as well as provide empirical support for identifying the community's most promising approaches. By promoting vigorous discussion of evaluation issues such as experimental methodologies, automated assessments, and shared tasks, the intelligent narrative technologies community can continue to grow and develop principled approaches for assessing and improving computational models of narrative.

Acknowledgements

The authors would like to thank the other members of the IntelliMedia Center for Intelligent Systems at North Carolina State University for useful discussions and support. This research was supported by the National Science Foundation under Grants REC-0632450, IIS-0757535, DRL-0822200 and IIS-0812291. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Aylett, R. and Louchart, S. 2008. If I were you: Double appraisal in affective agents. P*roc. of the 7th International Conference on AAMAS*, 1233-1236, Estoril, Portugal.

Bailey, P. 1999. Searching for storiness: Story-generation from a reader's perspective. *Working Notes of the AAAI Fall Symposium on Narrative Intelligence,* 157-163, Cape Cod, MA.

Belz, A., and Reiter, E. 2006. Comparing automatic and human evaluation of NLG systems. *Proc. of the 11th Conf. of EACL*, 313-320, Trento, Italy.

Callaway, C. and Lester, J. 2001. Evaluating the effects of natural language generation techniques on reader satisfaction. *Proc. of the 23rd Annual Conference of the Cognitive Science Society*, 164-169, Edinburgh, UK.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.) 2000. *Embodied Conversational Agents*. Boston, MA.: MIT Press.

Cavazza, M., Charles, F. and Mead, S.J., 2002. Planning characters' behaviour in interactive storytelling. *Journal of Visualization and Computer Animation 13*: 121-131.

Dix, A., Finlay, J., Abowd, G., and Beale, R. 2004. *Human-Computer Interaction*. Harlow, England: Pearson Education, Inc.

Dow, S., Mehta, M., Harmon, E., MacIntyre, B., and Mateas, M. 2007. Presence and engagement in an interactive drama. *Proc. of CHI*, 1475-1484, San Jose, CA.

Ekman, P. 2003. *Emotions Revealed*. New York: Henry Holt.

Gerrig, R. 1993. *Experiencing Narrative Worlds: On the Psychological Activities of Reading*. New Haven: Yale University Press.

Graesser, A.C., Lang, K.L., and Roberts, R.M. 1991. Question answering in the context of stories. *Journal of Experimental Psychology: General*, 120(3), 254-277.

Green, M., and Brock, T. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79(5),* 701-721.

Insko, B. E. 2003. Measuring presence: Subjective, behavioral and physiological methods. In G. Riva, F. Davide, & W. A. IJsselsteijn (Eds.), *Being There: Concepts, Effects and Measurements of User Presence in Synthetic Environments*. Amsterdam: IOS Press. 109-119.

Johnson, W.L. and Beal, C. 2005. Iterative evaluation of a large-scale, intelligent game for language learning. *Proc. of the 12th Intl Conference on Artificial Intelligence in Education*, 290-297, Amsterdam, The Netherlands.

Jurafsky, D. and Martin, J. 2008. *Speech and Language Processing*. Upper Saddle River, NJ.: Pearson Education, Inc.

Kelso, M., Weyhrauch, P., and Bates, J. 1993. Dramatic presence. *Presence: The Journal of Teleoperators and Virtual Environments*, 2(1), 1-15.

Lee, S., McQuiggan, S. and Lester, J. 2007. Inducing user affect recognition models for task-oriented environments. *Proc. of the 11th Intl. Conf. on User Modeling*, 380-384, Corfu, Greece.

Magerko, B. 2007. Evaluating Preemptive Story Direction in the Interactive Drama Architecture. *Journal of Game Development*, 2(3).

Mateas, M. and Stern, A. 2005. Structuring content in the Façade interactive drama architecture. *Proc. of AIIDE,* 93-98, Marina del Rey, CA.

McQuiggan, S., Robison, J., and Lester, J. 2008. Affective transitions in narrative-centered learning environments. *Proc of the 9th Intl Conf on ITS*, 490-499, Montreal, CAN.

McQuiggan, S., Rowe, J. and Lester, J. 2008. The effects of empathetic virtual characters on presence in narrative-centered learning environments. *Proc. of CHI*, 1511-1520, Florence, Italy.

McQuiggan, S., Rowe, J., Lee, S. and Lester, J. 2008. Story-based learning: the impact of narrative on learning experiences and outcomes. *Proc. of the 9th Intl Conference on ITSs*, 530-539, Montreal, Canada.

Mehta, M., Dow, S., Mateas, M. and MacIntyre, B. 2007. Evaluating a conversation-centered interactive drama. *Proc. of the 6th Intl Conf on AAMAS*, 1-8, Honolulu, HI.

Mott, B. and Lester, J. 2006. U-Director: A decision-theoretic narrative planning architecture for storytelling environments. In *Proc. of the 5th Intl Conf on AAMAS*, 977-984, Hakodate, Japan.

Nelson, M., Mateas, M., Roberts, D., and Isbell, C. 2006. Declarative optimization-based drama management in the interactive fiction Anchorhead. *IEEE Computer Graphics and Applications*, 26(3): 32-41.

Prince, G. 2003. *Dictionary of Narratology (Revised Edition)*. Lincoln, NE.: University of Nebraska Press.

Riedl, M. and Young, R. M. 2005. An objective character believability evaluation procedure for multi-agent story generation systems. *Proc. of IVA*, 278-291, Kos, Greece.

Riedl, M., Saretto, C. J., and Young, R.M. 2003. Managing interaction between users and agents in a multi-agent storytelling environment. *Proc. of the 2nd Intl Conf. on AAMAS*, 741-748, Melbourne, Australia.

Roberts, D., Nelson, M., Isbell, C., Mateas, M., and Littman, M. 2006. Targeting specific distributions of trajectories in MDPs. *Proc. of the 21st AAAI*, Boston, MA.

Rowe, J., McQuiggan, S., and Lester, J. 2007. Narrative presence in intelligent learning environments. *Working Notes of the 2007 AAAI Fall Symposium on Intelligent Narrative Technologies*, 126-133, Washington D.C.

Russell, S., and Norvig, P. 2003. *Artificial Intelligence – A Modern Approach*. Upper Saddle River, NJ.: Pearson Education, Inc.

Schubert, T., Friedmann, F., and Regenbrecht, H. 1999. Embodied presence in virtual environments. In Ray Paton & Irene Neilson (Eds.), *Visual Representations and Interpretations*. London: Springer. 269-278.

Si, M., Marsella, S.C., and Pynadath, D.V. 2005. THESPIAN: An architecture for interactive pedagogical drama. *Proc. of the 12th Intl Conf. on Artificial Intelligence in Education*, 21-28, Amsterdam, The Netherlands.

Turner, S. 1994. *The Creative Process: A Computer Model of Storytelling and Creativity*. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Walker, M., Litman, D., Kamm, C., and Abella, A. 1997. PARADISE: A framework for evaluating spoken dialogue agents. *Proc. of ACL*, 271-280, Madrid, Spain.

Weyhrauch, P. 1997. Guiding interactive drama. Ph.D. diss., Dept. of Computer Science, Carnegie Mellon Univ., Pittsburgh, PA.

Witmer, B. and Singer, M. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3), 225-240.

Zoll, C., Enz, S., Schaub, H., Aylett, R., and Paiva, A. 2006. Fighting bullying with the help of autonomous agents in a virtual school environment. *Proc. of the 7th Intl Conf. on Cognitive Modeling*, Trieste, Italy.