An Objective Character Believability Evaluation Procedure for Multi-Agent Story Generation Systems

Mark O. Riedl[1] and R. Michael Young[2]

[1] Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Los Angeles, CA 90292 USA
riedl@ict.usc.edu
[2] Department of Computer Science, North Carolina State University
Raleigh, NC 27695 USA
young@csc.ncsu.edu

Abstract. The ability to generate narrative is of importance to computer systems that wish to use story effectively for entertainment, training, or education. One of the focuses of intelligent virtual agent research in general and story generation research in particular is how to make agents/characters more lifelike and compelling. However, one question that invariably comes up is: Is the generated story good? An easier question to tackle is whether a reader/viewer of a generated story perceives certain essential attributes such as causal coherence and character believability. Character believability is the perception that story world characters are acting according to their own beliefs, desires, and intentions. We present a novel procedure for objectively evaluating stories generated for multiple agents/characters with regard to character intentionality – an important aspect of character believability. The process transforms generated stories into a standardized model of story comprehension and then indirectly compares that representation to reader/viewer mental perceptions about the story. The procedure is illustrated by evaluating a narrative planning system, Fabulist.

1 Introduction

Narrative as entertainment, in the form of oral, written, or visual stories, plays a central role in our social and leisure lives. Narrative is also used in education and training contexts to motivate and illustrate. The prevalence of narrative in our lives is partly due to what is called *narrative intelligence* which refers to the ability – human or computer – to organize experiences into narrative. Computational systems that reason about narrative intelligence are able to interact with human users in a natural way because they understand collaborative contexts as emerging narrative and are able to express themselves through storytelling. The standard approach to incorporating storytelling into a computer system, however, is to script a story at design time. That is, the system designers determine ahead of time what the story should be and hard-code the story into the system. An alternative approach is to generate stories either dynamically or on a per-session basis (one story per time the system is engaged

by a user). A system that can generate stories is capable of adapting stories to the user's preferences and abilities, has expanded "replay value" and is capable of interacting with the user in ways that were not initially envisioned by the system designers.

A story generation system is any computer application that creates a written, spoken, or visual presentation of a story – a sequence of actions performed by multiple characters. There have been many approaches to generating story: autonomous agents (e.g. [1] and [2; 3]); authorial planning (e.g. [4] and [5; 6]); models of creativity (e.g. [7]); models of dramatic tension (e.g. [8]); reactive selection of scene-like elements (e.g. [9; 10] to the extent that the drama will continue without active user participation). In some cases, the story emerges from real-time interaction between agents/characters. In other cases, a story is deliberately laid out by a single authoring agent and presented visually or as natural language. Regardless, one major drive of intelligent agent research is making agents – characters – more lifelike and believable.

For storytelling to be successful – to have an emotional or educational impact on the audience – a story must (a) be understandable and (b) believable in the sense that the audience is willing to suspend their disbelief. We argue that one property of story that affects both is *character believability*. Character believability refers to the numerous elements that allow a character to achieve the "illusion of life," including but limited to personality, emotion, intentionality, and physiology and physiological movement [11]. One important aspect of character believability is *character intentionality*. Character intentionality refers to the way in which the choice of actions and behaviors that a character makes appears natural (and possibly rational) to external observers. Character intentionality addresses the relationship of actions and behaviors to an agent's beliefs, desires, intentions as well as internal and external motivation.

The technical approach to automated story generation has implications for character believability and story coherence. There is a continuum between a strong autonomy approach and a strong story approach [9]. The strong story approach advocates centralized control of character behaviors. In general, centralized control, in the form of a single authoring agent that decides on the actions for all story world characters, is advantageous because the authoring agent can approach the story from a global perspective, choosing character actions in such a way that causal relationships are established [6]. Central control can be advantageous for character believability as well, helping to coordinate character actions in a way that eliminates the appearance of "schizophrenia" [12]. However, the more centralized control of character behaviors, the more likely it is that the characters will not be perceived by the reader/viewer as acting upon their own beliefs, desired, and intentions. This is particularly true of plan-based automated story generation systems in which the story planner is primarily concerned with establishing causal coherence of the story structure. In this case, character actions are chosen because of the effects they achieve and not necessarily based on whether it is believable for a character to perform an action.

Once the capability for story generation exists and stories are generated, evaluation becomes important. Evaluation of stories created by automated story generation systems often relies on subjective assessment. However, subjective assessment can

Once there was a Czar who had three lovely daughters. One day the three daughters went walking in the woods. They were enjoying themselves so much that they forgot the time and stayed too long. A dragon kidnapped the three daughters. As they were being dragged off, they cried for help. Three heroes heard the cries and set off to rescue the daughters. The heroes came and fought the dragon and rescued the maidens. Then the heroes returned the daughters to their palace. When the Czar heard of the rescue, he rewarded the heroes.

Fig. 1. An example story [14]**.**

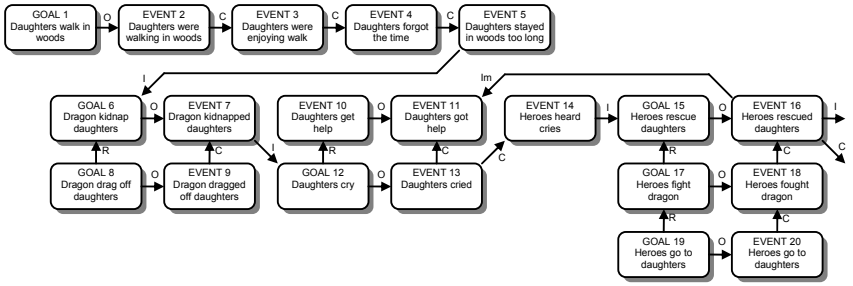

Fig. 2. A portion of the QUEST model for the story in Figure 1 [14].

be tangled up in many factors such as quality of natural language, subject interest in the topic of the story, novelty, and so on. An objective evaluation can be performed based on metrics such as story length or complexity of character roles (e.g. [3]) under the assumption that these metrics correlate to better stories. Other metrics may exist as well. Narrative has an impact on an audience; we believe that to understand the success of a narrative on an audience, one must measure the degree to which the reader/viewer perceives character intentionality since character is an integral part of story. A reader/viewer forms a mental model of the narrative structure and the characters in the narrative over time as the narrative unfolds and queries that mental model in order to actively predict outcomes and rationalize character actions [13]. The QUEST model of question-answering in the context of stories [14] provides a technique for devising a model of story comprehension for a story that can be indirectly compared to the mental model held in a subject's mind.

2 The QUEST Model of Question-Answering

The QUEST model [14] accounts for the goodness-of-answer (GOA) judgments for questions asked about passages of prose. One application of the QUEST model is to show that people build cognitive representations of stories they read that capture certain relationships between events in a story and the perceived goals of the characters in the story [14]. A reader's cognitive representation of the story is queried when the reader answers questions about the story. The types of questions supported by the QUEST model are: why, how, when, enablement, and consequence. For example, the story in Figure 1 has the corresponding QUEST knowledge structure shown in Figure 2. There are two types of nodes in the QUEST knowledge structure: event nodes, which correspond to occurrences in the story world, and goal nodes, which

correspond to goals that characters have. The links between nodes capture the different types of relationships between events and character goals.

- Consequence (C): The terminal event node is a consequence of the initiating event node.
- Reason (R): The initiating goal node is the reason for the terminal event node.
- Initiate (I): The initiating event node initiates the terminal goal node.
- Outcome (O): The terminal event node is the outcome of the initiating goal node.
- Implies (Im): The initiating event node implies the terminal event node.

Graesser et al. illustrate the QUEST model of question answering with the following question pertaining to the story in Figure 1: "Why did the daughters stay in the woods too long" (node 5)? There are many possible answers, some of which are:

A. Because the daughters forgot the time (node 4).
B. Because the dragon kidnapped the daughters (node 7).
C. Because the daughters were walking in the woods (node 2).
D. Because the heroes fought the dragon (node 18).

Both the question and each possible answer correspond to nodes in the knowledge structure. The QUEST model defines arc search procedures for each type of question (e.g. why, how, when, enablement, and consequence). The arc search procedures, starting at the queried node, distinguish between legal answer nodes and illegal answer nodes. That is, only nodes reachable by the arc search procedures are legal answer nodes. Answers (A) and (C) are legal answers. Of those two, (A) is preferred by the QUEST model because the corresponding node has a smaller structural distance from the queried node. The legality of answers and the weight of structural distance correspond to GOA judgments of human story readers.

3 Evaluation Procedure

To evaluate the character believability of a story generation system, we describe a procedure involving two conditions: a control condition and a test condition. The assumption is that a story generation system has been augmented or improved with regard to character believability. The control condition consists of a story generated by the story generation system without enhancement and/or augmentation while the test condition consists of a story generated by the same story generation system, but with enhancement. Given that the QUEST model of question-answering in the context of stories is empirically validated and that human narrative intelligence is relatively the same across subjects, the procedure for evaluating story generation systems is to compare an instance of a QUEST model of a specific generated story to subject comprehension of narrative structure for that story. In QUEST, "why" questions inquire about character goals, intentions, and motivations. The general idea behind the process described here is that a better story generation system (presumably the test condition) will result in stories whose structures better support human perception of character intentionality. The better the structure of the generated story, the better a QUEST representation of that story will predict reader/viewer question-answering. The procedure is as follows:

1. Generate control and test condition stories. Given two versions of a story generation system, generate a story from each. The two stories should be as similar as possible for evaluation to be possible. We assume that the story generation system will produce similar stories if given nearly identical initialization parameters but the test condition story will have elements and/or structure in the story that the control condition does not. Therefore, if there is a significant increase in the measure of understanding of character intentionality in the final results, then the enhancement to the story generation system does in fact improve the perception of character believability. Ideally, this will be achieved by initializing both systems with identical or nearly identical input parameters so as to avoid experimenter bias. It may not always be possible to use identical input parameters if the internal knowledge representations between versions of story generator are significantly different. In this case, one must control for the possibility that improvements are gained through different or increased knowledge.

2. Generate QUEST Knowledge Structures for each story. A QUEST Knowledge Structure (QKS) is an instantiation of a QUEST model for a particular story. This can be accomplished by hand or automatically. If done by hand, experimenter bias must be controlled for. To automatically generate a QKS from a story structure generated by a story generation system, there must be some formalized relationship between the data structures output by the story generation system and QUEST knowledge structures in general. For example, the results of [15] indicate a significant correlation between causal dependency plans and QKSs and validate the correlation experimentally.

3. Generate question-answer pairs. For each QKS, question-answer pairs can be composed from every possible combination of nodes in the QKS, as in [14]. It is important to compose both reasonable and nonsensical question-answer pairs. The study should focus on the "why" question-answer pairs since "why" questions emphasize understanding about intentional character actions. An example of a question-answer pair that can be generated from the QKS in Figure 2 is:

Q: Why did the heroes go to the daughters and dragon (node 20)?

A: Because the heroes heard the daughters cry (node 14).

Each question-answer pair will be rated by a subject on a Likert-type scale. The fact that there will be many question-answer pairs that occur in only one condition is not important since the analysis determines the degree to which subjects' mental models match a QKS within one condition and then compares that aggregate measure to the same from the other condition.

4. Use QUEST to identify "good" and "poor" question-answer pairs. QUEST specifies legal graph traversal routines which can be used to identify legal answers to questions. That is, if a legal graph traversal starting at the question node can find the answer node for a question-answer pair, then the question-answer pair is "good." This is a rough prediction of whether a subject's goodness-of-answer (GOA) rating of the question-answer pair will be favorable or not. For "why" questions, the arc search procedure searches for answer nodes by following forward reason arcs, backward initiate arcs, and backward outcome arcs [14]. The assumption is that the better a generated story supports human perception of character intentionality, the better the QKS for the story will predict subject GOA ratings.

5. Run subjects. Split subjects equally between the control and test conditions. The subjects make GOA judgments for the question-answer pairs, determining whether the answer seems like a reasonable response to the question. For each question-answer pair, there should be a Likert-type scale. For example, a four-point Likert-type scale has "Very bad answer", "Somewhat bad answer", "Somewhat good answer", and "Very good answer". Note that leaving out a middle ground ratings (e.g. "Neither agree nor disagree") forces a subject to commit to a positive or negative ranking, which is important because of the binary categorization of question-answer pairs. If the story is well-structured, subjects should rank question-answer pairs high when QUEST identifies the pairing as good and low when QUEST identifies the pairing as poor. Score the subject responses with numerical values. For example, "Very bad answer" gets a score of 1 and "Very good answer" gets a value of 4.

6. Compile and compare results. For each condition, find the between-subject mean for each question-answer pair. That is, the mean response value for question-answer pair X is *n*. For each condition, break the question-answer pairs into "good" and "poor" sets and find the mean response for each set. This gives you a 2-by-2 matrix of results: Mean response for "good" question-answer pairs versus mean response for "poor" question-answer pairs, and control condition versus test condition. For example, see Table 1. Favorable results are when:

- The mean GOA rating for "good" question-answer pairs for the test condition is statistically higher than the mean GOA rating for "good" question-answer pairs for the control condition.
- The mean GOA rating for "poor" question-answer pairs for the test condition is statistically lower than the mean GOA rating for "poor" question-answer pairs for the control condition.

4 Example – Story Planning

We illustrate our evaluation technique by evaluating previous research on a story generation system based on partial order planning. The story generation system is called Fabulist [6] and utilizes an Intent-driven Partial Order Causal Link (IPOCL) planner [5; 6] that is an enhancement of a more conventional Partial Order Causal Link (POCL) planner, specialized to narrative generation.

Young [16] suggests that planning has many benefits as a model of narrative. First of all, plans are comprised of partially ordered steps. If the plan steps represent actions that are performed by characters in the story world, then a plan makes a good model of a story fabula – the chronological enumeration of events that occur in the story world between the time the story begins and the time the story ends. Secondly, planners such as UCPOP [17] construct plans based on causal dependencies. Causal dependency planning ensures that all character actions are part of a causal chain of events that lead to the outcome of the story, resulting in a coherent story structure.

The causal dependencies between character actions and the story outcome ensure coherent story structure, but also pose a problem for character believability. Specifically, causal dependency planners attempt to find a sequence of operations that

achieve a particular goal. In the case of a story planner, character actions are not chosen because they are the natural (e.g. believable) thing for a character to do at a particular time, but because they establish causal relationships that are necessary for plan soundness. Conventional planners do not reason explicitly about character intentionality and, consequently, their story plans are not guaranteed to possess this property. For example, the Universe story generation system [4] uses a hierarchical planner to piece together plot fragments into a story plan. Plot fragments are decomposed into character actions. Universe, however, selects plot fragments (and consequently character actions) to be in the story plan only when they establish causal conditions necessary to achieve the story outcome.

Our enhanced narrative planner, IPOCL, reasons about possible character intentions in order to construct narrative plans that not only have causal coherence but also motivate the actions that story world characters have. The hypothesis is that our narrative planner will generate better structured narratives that facilitate reader/viewer perception of character believability (or at least character intentionality).

4.1 Fabulist

Fabulist [6] is a story generation system that uses a causal dependency planner to create a story involving multiple characters that are possibly antagonistic toward each other. The causal dependency planner accepts a description of the initial state of the story world, a partial description of the outcome that should result from the events of the story, and a library of actions that characters can perform. The output of the planner is a story plan where the operations of the plan are actions performed by story world characters.

The causal dependency planner used by Fabulist is a special planner designed for story generation called the Intent-driven Partial Order Causal Link (IPOCL) planner [5; 6] (although in the control condition of our evaluation, a conventional POCL planner will take its place for comparison purposes). In addition to the narrative planner, Fabulist also has a discourse planner and a media realizer that are configured in a pipeline. The narrative planner generates a narrative plan which describes all the events that will happen in the story world between the time the story begins and the time the story ends. The discourse planner takes the narrative plan as input and generates a narration of the story. The discourse plan consists of the communicative actions required to tell the story to an audience. Fabulist uses an unmodified version of the Longbow discourse planner [18]. The media realizer takes the discourse plan as input and generates natural language. Fabulist currently uses a simple template-matching routine to generate surface-level text, although a more sophisticated system such as that in [19] could be used instead.

The IPOCL planning algorithm addresses the limitations of conventional causal dependency planners when applied to story generation. Specifically, conventional planners make certain assumptions that are not valid in the domain of story planning.

– The planner is creating a plan for a single agent.
– The goal of the planning problem is the desired world state of the agent.

In contrast, a single planner that is creating a story plan must create a plan for multiple agents (story world characters). In addition, the goal of the planning problem represents the *outcome* of the story as intended by the human author. That is, the outcome is not necessarily intended by any character and most likely not intended by all characters. If the outcome is intended by all story world characters, then the characters will appear to collaborate to bring about the outcome. However, it is more likely that many characters do not share the same goals and may even have conflicting goals.

The IPOCL algorithm addresses the mismatch between conventional planning and story planning by decoupling the characters' intentions from the author's intentions. IPOCL does not assume that the story world characters intend the outcome (goal state) of the story. Instead, IPOCL (1) searches for the intentions that each character might have and (2) motivates through story events why those characters have the intentions that they do. At stake is the perception that a character has goals, that those goals are formed in reaction to stimuli, and that the character is acting to achieve those goals.

IPOCL is based on conventional causal dependency planners such as UCPOP [17]. However, IPOCL story plans contain richer structural representation because it includes character intentions that are distinct from the story goal and, consequently, tend to be longer than conventional plans. That is, given the same initialization parameters, IPOCL and a conventional planner would generate different plans. But does the IPOCL story plan support audience perception of character intentionality better than one generated by a conventional planner? We apply our objective evaluation procedure to determine this.

4.2 Method

To determine whether subjects perceived character intentionality in stories generated by Fabulist, we used two versions of Fabulist to generate two similar narratives. Subjects were separated randomly into groups, asked to read one of the stories, and rate the goodness of answer of question/answer pairs relating to the story they read. One version of Fabulist had a story planner component implementing the IPOCL algorithm, while the other used a conventional causal dependency planner. Both versions of Fabulist had identical discourse planner components based on the Longbow planner [18], and identical template-based text realizer components. Both versions of Fabulist were initialized with identical parameters.

A QUEST knowledge structure (QKS) – a representation of the cognitive structures held in the mind of a reader of a story – is a directed acyclic graph of events and goals. As such, QKSs are similar to plans, which are also directed acyclic graphs of events and goals. Christian and Young [15] define a procedure by which a simple yet functional QKS can be derived from a plan. They demonstrate that the QKS generated from a plan significantly predicts the goodness-of-answer judgments for "why"

There is a woman named Jasmine. There is a king named Mamoud. This is a story about how King Mamoud becomes married to Jasmine. There is a magic genie. This is also a story about how the genie dies.

There is a magic lamp. There is a dragon. The dragon has the magic lamp. The genie is confined within the magic lamp.

There is a brave knight named Aladdin. Aladdin travels from the castle to the mountains. Aladdin slays the dragon. The dragon is dead. Aladdin takes the magic lamp from the dead body of the dragon. Aladdin travels from the mountains to the castle. Aladdin hands the magic lamp to King Mamoud. The genie is in the magic lamp. King Mamoud rubs the magic lamp and summons the genie out of it. The genie is not confined within the magic lamp. The genie casts a spell on Jasmine making her fall in love with King Mamoud. Jasmine is madly in love with King Mamoud. Aladdin slays the genie. King Mamoud is not married. Jasmine is very beautiful. King Mamoud sees Jasmine and instantly falls in love with her. King Mamoud and Jasmine wed in an extravagant ceremony.

The genie is dead. King Mamoud and Jasmine are married. The end.

Fig. 3. Text of story in control condition.

There is a woman named Jasmine. There is a king named Mamoud. This is a story about how King Mamoud becomes married to Jasmine. There is a magic genie. This is also a story about how the genie dies.

There is a magic lamp. There is a dragon. The dragon has the magic lamp. The genie is confined within the magic lamp.

King Mamoud is not married. Jasmine is very beautiful. King Mamoud sees Jasmine and instantly falls in love with her. King Mamoud wants to marry Jasmine. There is a brave knight named Aladdin. Aladdin is loyal to the death to King Mamoud. King Mamoud orders Aladdin to get the magic lamp for him. Aladdin wants King Mamoud to have the magic lamp. Aladdin travels from the castle to the mountains. Aladdin slays the dragon. The dragon is dead. Aladdin takes the magic lamp from the dead body of the dragon. Aladdin travels from the mountains to the castle. Aladdin hands the magic lamp to King Mamoud. The genie is in the magic lamp. King Mamoud rubs the magic lamp and summons the genie out of it. The genie is not confined within the magic lamp. King Mamoud controls the genie with the magic lamp. King Mamoud uses the magic lamp to command the genie to make Jasmine love him. The genie wants Jasmine to be in love with King Mamoud. The genie casts a spell on Jasmine making her fall in love with King Mamoud. Jasmine is madly in love with King Mamoud. Jasmine wants to marry King Mamoud. The genie has a frightening appearance. The genie appears threatening to Aladdin. Aladdin wants the genie to die. Aladdin slays the genie. King Mamoud and Jasmine wed in an extravagant ceremony.

The genie is dead. King Mamoud and Jasmine are married. The end.

Fig. 4. Text of story in test condition.

and "how" questions when arc search procedure was considered without structural distance[1].

Both the test condition story and the control condition story are generated from the same set of inputs. The stories differ due to the fact that the test condition story planner reasons about character intentions distinct from the outcome and introduces additional motivating actions into the story to provide explanation for why characters act. The story in the control condition has 10 events and is shown in Figure 3, while the story in the test condition has 13 events and is shown in Figure 4. Figures 5 and 6 show QKS representations of the control condition story and test condition story, respectively. The narrative plans from which the QKSs are derived are not shown here; see [6] for more details. Note that there are significant similarities between the

[1] An additional study by the authors (not reported) determined that QKSs derived from IPOCL plans significantly predict GOA judgments when structural distance is ignored ($p < 0.0005$).

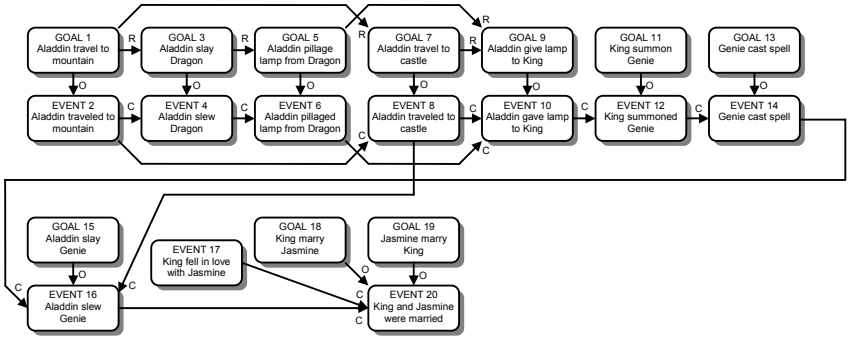

Fig. 5. QKS for the story in the control condition.

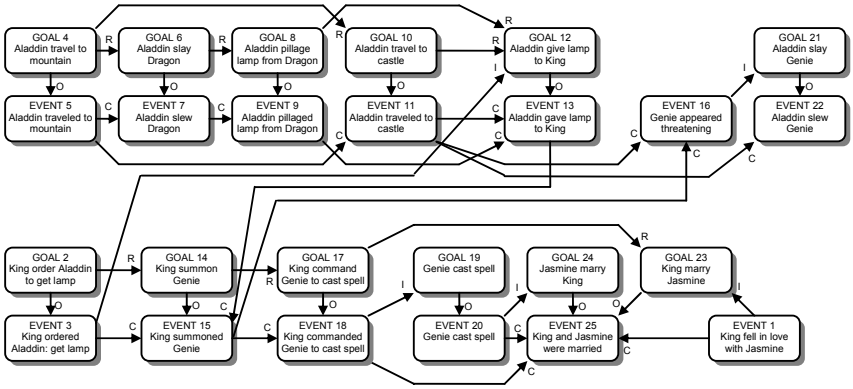

Fig. 6. QKS for the story in the test condition.

two stories, making a comparison study possible. Specifically, the set of events in the test condition story is a superset of the events in the control condition story.

There is one distinct action ordering difference between the two fabula plans: the event where the King falls in love with Jasmine is temporally constrained to occur first in the test condition story but is under-constrained in the control condition story and happens to fall late in the story. Had it come earlier in the control condition, some subjects may have *inferred* a relationship between the king falling in love and Aladdin's actions even though there is no actual relationship in the QKS. However, the ordering of this particular action does not affect the QKS representation because QUEST does not capture the temporal relationship between events beyond what is needed for causal coherence. From this, we conclude that the ordering will only have an insignificant impact on the results of the comparison between subjects' mental models and the QKS for the story.

The evaluation was set up as a questionnaire in which subjects read a story and then make goodness-of-answer (GOA) judgments about pairs of question and answers. A question-answer pair has a "why" question about an intentional action performed by a character in the story and a possible answer. For example, the question, "Why did Aladdin slay the dragon?" might be paired with the answer, "Because King

Mamoud ordered Aladdin to get the magic lamp for him." The subjects were asked to rate the goodness of the answer for the given question on a four-point Likert scale ranging from "Very bad answer" to "Very good answer." The subjects were shown examples of a question-answer pairs before the rating task began, but were not otherwise given a definition of "good" or "poor" or trained to make the judgment. Subjects rated the GOA of a question-answer pair for every combination of goal nodes in a QKS for the story. Subjects were asked to read the story text completely at least once before proceeding to the ratings task and were allowed to refer back to the original text at any time during the rating task. The control condition questionnaire had 52 question-answer pairs while the test condition questionnaire had 82 question-answer pairs due to the increased story plan length. The question-answer pairs in each condition were evaluated by QUEST. "Why" questions were classified as "good" or "poor" based on the arc search procedure following forward reason arcs, backward initiate arcs, and backward outcome arcs [14] applied to the QKS derived from the story plan for the particular condition. The aim was to determine if there was a statistically significant difference in subjects' mean agreement with the relevant QKS between conditions. An example of a question-answer pair that is likely to be judged as "good" in the test condition but judged ambiguously in the control condition is:

Q: Why did Aladdin travel from the castle to the mountains?

A: Because King Mamoud wanted to rub the magic lamp and summon the genie.

In the test condition story, the story explicitly motivates Aladdin's sequence of actions involving traveling into the mountains and slaying the dragon – Aladdin is ordered to get the King the magic lamp. In the control condition the reason for Aladdin's sequence is left unmotivated and some readers will infer the answer to be a justifiable reason (especially in hindsight) while others will not.

Thirty-two undergraduate students in the Computer Science program at North Carolina State University participated in the study. All subjects were enrolled in the course, *Game Design and Development*, and were compensated for their time with five extra credit points on their final grade in the course.

4.3 Results

Each question-answer pair in each questionnaire was assigned a "good" rating or a "poor" rating based on the QUEST prediction. The results of subjects' answers to questionnaire answers are compiled into Table 1. The numbers are the mean GOA ratings for each category and each condition. The numbers in parentheses are standard deviations for the results.

A standard one-tailed t-test was used to compare the mean GOA rating of "good" question-answer pairs in the test condition to the mean GOA rating of "good"

Table 1. Results for character intentionality evaluation.

	Mean GOA for "good" Q/A pairs (std. dev.)	Mean GOA for "poor" Q/A pairs (std. dev.)
Test condition	3.1976 (0.1741)	1.1898 (0.1406)
Control condition	2.9912 (0.4587)	1.269 (0.1802)

question-answer pairs in the control condition. The result of the t-test with 15 degrees of freedom yields $t = 1.6827$ ($p < 0.0585$). Subjects in the test condition had significantly higher GOA ratings for "good" question-answer pairs than subjects in the control condition.

A standard one-tailed t-test was used to compare the mean GOA rating of "poor" question-answer pairs in the test condition to the mean GOA rating of "poor" question-answer pairs in the control condition. The result of the t-test with 15 degrees of freedom yields $t = 1.8743$ ($p < 0.05$). Subjects in the test condition had significantly lower GOA ratings for "poor" question-answer pairs than subjects in the control condition.

Favorable results were achieved for each relevant comparison. From this we can conclude that the story in the test condition supported reader comprehension of character intentionality better than the story in the control condition. It is reasonable to infer that the improvement of the test condition over the control condition is due to enhancements to the automated story generation capability.

4.4 Discussion

There is a large degree of commonality between the two stories generated in the study, suggesting that the additional content in the IPOCL (test condition) plan had an impact on subject comprehension of character intentionality. Since subjects in the test condition are more in agreement with the QUEST model than subjects in the control condition, we conclude that stories generated by a story planner implementing the IPOCL planning algorithm support a reader's comprehension of character intentionality better than stories generated by a story planner implementing a conventional POCL planner. However, there were limitations to our study that must be taken into consideration. These limitations are largely due to our use of a novel evaluation technique and consequently the inability to foresee difficulties. We present them here as lessons learned during the application of the evaluation methodology.

The standard deviation for the control condition and "good" question-answer pairs was high. Further analysis reveals that subjects are likely to judge a question-answer pair as "good" if there is lack of evidence against the possibility that the character action might have been intentional. We speculate that reader/viewers simultaneously consider multiple hypotheses explaining character behavior until they are disproved. Regardless of the content of any communicative act, one will always be able to provide a more or less plausible explanation of the meaning [20].

One independent variable we failed to control for was story length and complexity. It is possible that the effects we measured were a result of story length and complexity instead of improved story structure generated by the story generation system. We believe this to be unlikely, but future evaluations should add to the control condition story hand-written filler sentences that do not impact character believability so that it matches the length and complexity of the test condition.

A second limitation to the evaluation, as we have already noted, was the lack of control for partial ordering of actions in the control condition. Since the story planners used for the evaluation were least-commitment planners, they did not commit to

a total ordering of actions unless necessary. A total order was artificially imposed on partially-ordered action sequences so that the plans could be rendered into natural language. To be thorough we would have had to consider different total orderings to determine if ordering had an effect on reader comprehension of character intentionality. The QUEST model remains the same for all possible, legal orderings since it factors out temporal considerations that are not relevant to causality. This leads us to conclude that different orderings would not significantly impact our results. In fact, having the King fall in love with the princess sooner will likely have resulted in a wider range of GOA judgments to some question-answer pairs, making the standard deviation in the control condition higher and the difference in means with the test condition larger.

A final limitation to our evaluation of Fabulist is related to our simplistic domain modeling of discourse generation. The Longbow discourse planner [18] is a very powerful tool for discourse generation. However, we used a simplified model of discourse structures that caused explicit statements of character intention to be rendered into the story text for the test condition. That is, subjects in the test condition were told how the characters formed their intentions. We believe that our results would be the same if these explicit statements were excluded because human readers are very good at inferring intentions from observations of actions. However, to be complete, we would have to control for such artifacts from discourse generation.

5 Conclusions

The ability to computationally generate stories can result in computer systems that interact with humans in a more natural way. To date story generation systems have used autonomous multi-agent technologies and single authoring agent approaches. Regardless of the technology, automated story generation continues to improve, particularly within the bounds of character believability. It is useful, therefore, to be able to evaluate the degree to which enhancements to story generation technology improves the quality and character believability of generated stories. Instead of using subjective measures, we present a process for objectively assessing the degree of enhancement to character intentionality – one important aspect of character believability – in generated stories. The process relies on the fact that a reader/viewer's perception of character intentionality can be compared to a QUEST representation of the story because QUEST is a validated model of human question-answering in the context of stories. We present the evaluation process and illustrate it by describing how it was applied to the evaluation of the Fabulist story generation system.

Acknowledgements

This work has been supported by NSF CAREER award 0092586. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Meehan, J.R.: *The Metanovel: Writing Stories by Computer*. Ph.D. Dissertation, Yale University (1976).
2. Cavazza, M., Charles, F., &Mead, S.: Planning characters' behaviour in interactive storytelling. *Journal of Visualization and Computer Animation*, vol. 13 (2002) 121-131.
3. Charles, F. & Cavazza, M.: Exploring the scalability of character-based storytelling. *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi Agent Systems* (2004).
4. Lebowitz, M.: Story-telling as planning and learning. *Poetics*, vol. 14 (1985) 483-502.
5. Riedl, M.O. & Young, R.M.: An intent-driven planner for multi-agent story generation. *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi-Agent Systems* (2003).
6. Riedl, M.O.: *Narrative Planning: Balancing Plot and Character*. Ph.D. Dissertation. North Carolina State University (2004).
7. Turner, S.R.: *The Creative Process: A Computer Model of Storytelling*. Hillsdale, NJ: Lawrence Erlbaum Associates (1994).
8. Szilas, N.: IDtension: A narrative engine for interactive drama. *Proceedings of the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment* (2003).
9. Mateas, M.: *Interactive Art, Drama, and Artificial Intelligence*. Ph.D. Disssertation, Carnegie Mellon University (2002).
10. Mateas, M. & Stern, A.: Integrating plot, character, and natural language processing in the interactive drama Façade. *Proceedings of the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment* (2003).
11. Bates, J.: The role of emotion in believable agents. *Communications of the ACM*, vol. 37 (1994).
12. Sengers, P.: Narrative and schizophrenia in artificial agents. In M. Mateas and P. Sengers (Eds.) *Narrative Intelligence*. John Benjamins, Amsterdam (2003).
13. Gerrig, R.J.: *Experiencing Narrative Worlds: On the Psychological Activities of Reading*. Yale University Press, New Haven (1993).
14. Graesser, A.C., Lang, K.L., & Roberts, R.M.: Question answering in the context of stories. *Journal of Experimental Psychology: General*, vol. 120 (1991).
15. Christian, D.B. & Young, R.M.: Comparing cognitive and computational models of narrative structure. *Proceedings of the 19th National Conference on Artificial Intelligence* (2004).
16. Young, R.M.: Notes on the use of planning structures in the creation of interactive plot. In: M. Mateas and P. Sengers (Eds.): *Narrative Intelligence: Papers from the 1999 Fall Symposium*. American Association for Artificial Intelligence, Menlo Park CA (1999).
17. Penberthy, J.S. & Weld, D.: UCPOP: A sound, complete, partial-order planner for ADL. *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning* (1992).
18. Young, R.M., Moore, J.D., & Pollack, M.E.: Towards a principled representation of discourse plans. *Proceedings of the 16th Conference of the Cognitive Science Society* (1994).
19. Callaway, C.B. & Lester, J.C.: Narrative prose generation. *Artificial Intelligence*, vol. 139 (2002).
20. Sadock, J.M.: Comments on Vanderveken and on Cohen and Levesque. In: P.R. Cohen, J. Morgan, and M.E. Pollack (Eds.): *Intentions in Communication*. The MIT Press, Cambridge MA (1990) 257-270.