Towards Automated Story Analysis Using Participatory Design

ABSTRACT

Involving school teachers in the development of the intelligent writing tutor StoryStation allowed progress to be made on the problem of story understanding in artificial intelligence. An experienced Scottish school-teacher developed a rating scale and guidelines for StoryStation's automated plot analysis agent in the story rewriting task. In this task, pupils rewrite a story in their own words, allowing them to devote their full attention to improving their writing technique instead inventing a new plot. If the pupil forgets or confuses significant parts of the plot, or entirely forgets the story, the software may alert them or their teacher. Teacher participation in the creation of the rating scale guided both the development of computational linguistic tools used to analyze the stories, and teachers and children helped shape the range and scope of the plot analysis agent. A teacher and a story-teller rated the corpus, and this scale was used to successfully train the agent to identify both "good" and "poor" stories. Identification of "excellent" and "fair" stories proved to be very difficult. Upon reflection with teachers and a comparison of the methodology used by our agent with discourse psychology, a number of facets of story understanding are shown to be beyond the range of the automated plot analysis agent, and the advantages and disadvantages of automated plot analysis are weighed, including social factors.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

PDC Proceedings, story understanding, participatory design, plot analysis, computational linguistics, intelligent tutoring systems

1. INTRODUCTION

Participatory Design Conference '04 Toronto, Canada

StoryStation[23] is an intelligent tutoring system to aid children in developing their writing abilities, developed using a child-centered design methodology adapted from Druin[7] and Scaife[27]. In the tradition of Kyng[15], research was based on participatory design with teachers and students in Scottish primary schools[24]. In the course of design, animated agents were developed that help the child with spelling, diction, and character development. During the initial design of the system, eight students and two teachers were consulted, and further feedback by both teachers and students was elicited through experimental deployment of StoryStation in two Scottish schools as detailed by Roberston[25]. The agents in StoryStation also provide access to a dictionary, thesaurus, and a tour guide of the system. Assessment of spelling, vocabulary and characterization skills is also available. These agents are represented as animal characters that float on the screen of the pupil and offer advice and support, and their icons were created by children themselves. Instead of providing negative criticism, StoryStation praises the pupil for their good work, and provides positively-phrased constructive comments to help the pupils. One agent that was requested by both children and teachers, tentatively called "Pinky the Plot Analyzer", is currently not functioning.

The plot analysis agent is currently not working in StoryStation due to a lack of both a theory and implementation of automated plot analysis. Teachers would like this agent to be able to provide some type of score for the plot of the story, and then give advice to the student on whether or not they should work more on the plot of the story. Ideally, a more fine-grained level of analysis in which the agent reminds the pupil of particular pieces of the plot they may have left out (like missing characters and events) would be highly appreciated by both children and teachers. Some children become frustrated in the writing process when asked to think of a plot themselves, often leading to incomplete writing assignments. An automated plot analysis agent would provide encouragement for the child to continue writing the plot. The agent must be able to analyze the child's plot for missing or confused characters, episodes, and other events. The agent should remind the child if they have forgotten or misconstrued an event, and encourage the child to write more when they get stuck by recalling specific events for them. More importantly, the agent should be able to assess the general quality of the rewritten plot, giving the child general feedback, such as "Good job remembering the story!" or "Would you like for me to retell you the story?"

In this project, the plot analysis agent for the story rewrit-

ing task, a common writing task in Scottish primary schools, was designed by teachers in the Scottish school system together with researchers from the University of Edinburgh. The plot analysis agent uses techniques from computational linguistics, in particular extraction of events from the story and statistical rater modeling, in order to build a model of how a teacher actually assesses the plot of a story in the story rewriting task. To this end, a corpus was collected of stories rewritten by children, and these stories were rated by raters using a teacher-designed metric. The plot analysis agent, by relying on its ability to learn how humans perform a task, is able to perform reasonably well on a task thought to be classically very difficult for artificial intelligence. This agent then can automatically aid students in their recall of the plot, relieving tedium from the teachers. This allows teachers to concentrate their teaching on other aspects of writing rather than correcting mistakes in plot recall and development, and allows students to further hone their individual learning skills.

2. THE STORY REWRITING TASK

In the *story rewriting task*, children write a story they have heard before in their own words. The cognitive load of inventing an entirely new plot is taken off the students through the story rewriting task, allowing them to devote their full attention to their writing technique. A student could work on issues such as diction and spelling, instead of inventing a new plot structure. Teachers in our study had found this a very effective way of getting students in particular to describe a scene or character in depth.

A series of three story rewriting tasks were performed from classes at Methilhill and Torbain Primary Schools in Kirkcaldy, Scotland by Judy Roberston. The children, ages 10-12 and from a broad range of reading levels and socio-economic backgrounds, were told a story by a storyteller, called the *original story* throughout this paper. The children were asked to rewrite the story in their own words. A transcript of the story as told by the storyteller was collected, and the rewritten stories were transcribed. The stories were collected into a digital corpus of 103 stories.

For our corpus the story-teller told the students a story called "Nils' Adventure," a story from "The Wonderful Adventures of Nils" by Selma Lagerloff[17]. A transcript of the story as told by the storyteller is available[12]. The story involves a boy called Nils who jumps on the back of a talking stork, which drops him off on a beach where the boy finds an old, green coin. Thinking it useless, Nils throws the coin away. A city appears from the waves, and its residents offer Nils wonderful wears for the price of only one coin. Nils runs outside to retrieve the coin, and the city disappears. The stork explains to Nils that the city was cursed to appear only once every hundred years due to the greed of its inhabitants. Only buying something from them will dispel the curse. Disappointed by his lack of forethought, Nils and the stork leave for another adventure.

In summary, the approach taken in this study was to combine techniques from participatory design with computational linguistic tools to analyze children's stories to provide automated feedback which approximates feedback given by teachers for the story rewriting task. This task is singled out because teachers have suggested that automated plot analysis is the one feature of StoryStation they would most like to see working.

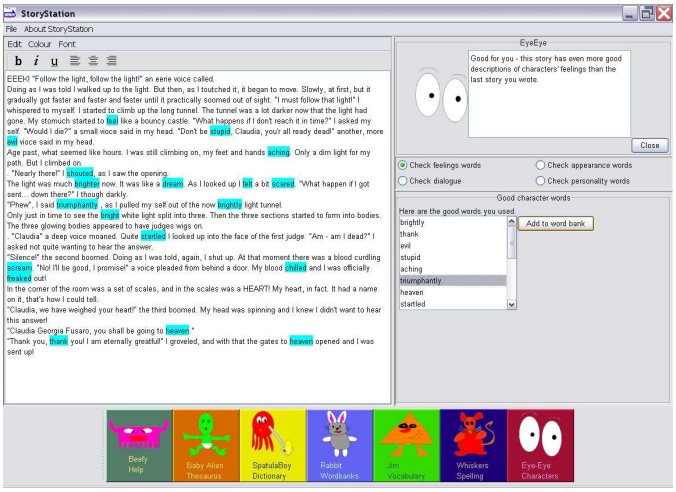

Figure 1: Graphical Interface to StoryStation

3. STORYSTATION: WORKING WITH PARTICIPANTS

StoryStation follows in tradition of intelligent tutoring systems outlined by Britton[3] that are informed by research on the cognitive psychology of writing. Flower[9] emphasized how writers must manage many differing constraints, and that students often need help on each of the constraints both individually and in tandem. This "constraint"-theory also mirrored the concerns and goals of StoryStation as voiced by the teachers. These constraints are visually manifested as animated agents with differing specialties that can be invoked by the child on demand. See *Figure 1* for a picture of the graphical layout of StoryStation and its agents.

The teachers and students in the design team did a comparison of other writing tutor software in the field, and helped chose the number and kinds of features the agents in StoryStation embodied, including automated plot analysis. Note that no current intelligent tutoring system besides the WRITE system by ETS employs automated plot analysis[4], but both students and teachers felt it would be very useful as regards recalling narrative texts. Students in the design team wanted the responses of the agents to be individually tailored towards their writing levels, so StoryStation keeps a student model of each student individually, allowing StoryStation to use their current curriculum level and the level which the child is working towards in shaping its agents behavior. If the student has never used StoryStation before, a default model for the student's current curriculum level is used.

Students who took part in a field study in a state funded primary school felt that the agents helped their writing. Comments included "It made me feel more confident. You know you're not making mistakes in words." and "It made me feel happy because it was helping me with my spelling and words I didn't know." As detailed by Robertson[25], students felt StoryStation complimented their teacher well, with 57% of students in a sixty student questionaiire responded that "They would be more likely to trust advice from StoryStation than a teacher," although many students said they would trust a teacher more since "a teacher knows

what your writing is meant to be about, but StoryStation doesn't"[25]. Students liked using StoryStation because "You don't get embarassed if your forget. [In class] you have to go up and ask again. So you keep on asking if you forget and sometimes the teacher shouts at you if get something wrong." Since the help from the agents in StoryStation is purely optional, one student said that if the StoryStation agents "yelled at him" he could "tell them to shut up." Many students appreciated the specific nature of the advice, and one student when comparing StoryStation to a teacher said, "The computer can be more specific the teacher just says just go and write more" and "The teacher might help you with vocabulary but she won't go through every single one [word]"[25]. StoryStation has also been evaluated with 18 teachers who found the system to be very promising, especially its potential to allow the students to develop their writing skills independently[23]. There is demand for intelligent tutoring systems to help children learn writing skills, and by involving teachers and students in the process using participatory design so far has resulted in a satisfactory system. The addition of a plot analysis agent that could "understand stories" would bring the system to its fullest potential.

4. THE STORY UNDERSTANDING PROBLEM

Story understanding has been studied in a formal framework since the dawn of cognitive science and artificial intelligence. This quintessential of human activities is central to our cultures and our ability to navigate in the world, and so any ability to mechanize such a process would give valuable insight into how the human mind works. Researchers from a wide variety of backgrounds, including anthropologists interested in folklore and computer scientists interested in natural language understanding and generation, have applied their wits to the problem. Although the work in formal analysis of stories goes back as far as the structural analysis Claude Levi-Strauss[19] and the functional analysis of Propp[22], the foundational work in story analysis comes from the seminal "Notes for a schema of stories," in which Rumelhart used a *story grammar*, a series of production rules, to describe the structure of stories as a context-free grammar[26]. He claims that "just as simple sentences can be said to have an internal structure, so too can stories be said to have an internal structure. This is in spite of the fact that no one has ever been able to specify a general structure for stories that distinguishes the strings of sentences which form stories from strings which do not"[26]. While the spirit of Rumelhart's analysis closely follows the classic production rules of context-free grammars for sentences, each story actually has a sentence as a terminal symbol with non-terminals being a functional grouping of the sentences. An example of a story grammar production rule is: Story $\Rightarrow$ Setting + Episode.

This approach is problematic. Story grammars are incapable of detecting the presence or absence of a particular character in an event. Sentences are limited to those which can be reduced to a single proposition and the number of categories are upon closer inspection ill-defined and arbitrary, making it impossible for either a computer algorithm or human teacher to use such a scheme in a teaching environment.

4.1 Computers and Story Understanding

While this previous approach was done manually by cognitive researchers, the first attempt to automate story understanding using computers was the doctoral dissertation of Eugene Charniak, "Towards a Model of Children's Story Comprehension," under the supervision of Marvin Minsky[5]. Understanding children's stories was viewed as a grand challenge in the heyday of artificial intelligence. If a computer could somehow capture the world-knowledge and cultural context needed to understand a story, it would be a significant victory for artificial intelligence. Children's stories were also focused upon as they were considered to be the simplest of stories for a computer to understand. Children's stories were viewed in a negative manner: If a computer could not understand "something as simple" as a children's story, then it could not be intelligent.

The problem of understanding children's stories was shown to be far more difficult than originally surmised. Stories take place in an immense cultural and "common-sense" context, and a computer knows none of it. Researchers attempted to hand-code this knowledge into *frames*, which represent cultural and world knowledge as a structure of named variable slots that can be given a set of values to represent a particular story[21]. These variables are often be determined by the operation of a logical inference engine. A narrative plot is a series of frames, with the content and explanation of a story determined by values given from a computational analysis of the sentences and the operation of an inference engine.

Research is still done on applying AI planning techniques to story understanding[6]. However, for the most part these efforts were one of the first grand failures of the classical artificial intelligence approach: the world and cultural knowledge needed to understand a story far outstrips our ability to even consciously iterate through such knowledge, much less formalize it. Sometimes inference is often either transparent (such as "Let's have an adventure") or simple ("If you want to fly, get on the back of the flying stork"). Logic itself is far from "common-sense" in most children's stories, as when Nils forgetting to return a coin causes the city to disappear under the waves. Do we demand a "fantastical-sense" inference engine, and what could that possibly entail? Must we demand not only the ability to predict the common-place, but the uncommon from artificial intelligence? In "Nils' Adventure", should our computer notice if the coin is left out by the pupil, and the city still disappears? A coin that causes a city to disappear will just not exist in common-sense databases of causal relations. Additionally, the whole frame approach can not be employed since it would be infeasible to ask a teacher to handcode each student's story into representation suitable for processing by a frame system. It is the extraction and labeling of events from the writing of the children that is of concern, not any system based on planning and logic.

4.2 The Frame Problem

The extraction of events from a story still involves using some formal model of the children's story that can be implemented on a computer. All formal models are haunted by what is termed the "Frame Problem." In its original formulation by McCarthy and Hayes[20], it was realized that most models are inherently static as regards time, and that the addition of time to a model can lead to consequences un-

predictable by the model itself. More generally, the Frame Problem consists of the modeler not knowing the unintended consequences of leaving *something* out of their model, so that the model does not do justice to the real world problem.

The Frame Problem can be rephrased as the Quantification Problem: How many and what things do we formalize into the model? It would be impossible for us to formalize everything. If we develop a model of the plot as a series of events, and create a model of the pupil's rewritten story as another series of events, and try to compare them event-by-event, our algorithm will probably fail to match many perfectly good rewritten stories unless the pupil exactly copies the story word for word. The pupil may add in a detail like "When the stork flew away, Nils left Sweden," which is completely true but not explicit in the original story. The pupil may leave out unimportant details due to their irrelevance to the flow of the story. In the introduction of "Nils' Adventure," Nils is found riding geese. Yet Nils riding on geese has little to do with the rest of the story. The model must clearly be *flexible*. We could let human experts formalize and quantify all possible events they deems relevant for a particular story. This would make our product less portable over new stories that a teacher wants to use, and would involve another layer of expertise outside the teacher. The alternative is to let the computer only use whatever events it can find automatically. Both options clearly are ridden with problems. The human expert in finding events may be error-prone, but the computer may miss events or fail to recognize events that help the story make more sense. Our story introduces Nils as a Swedish boy that flies on geese. Much more could have been said in the introduction to the story that may help make the events of the story make sense, such as mentioning Nils' former cruelty to animals that he is trying to atone for? Was it even really relevant for the geese to be mentioned, since for the rest of the story Nils interacts with a stork? These are hard judgments for a human to make, and no obvious algorithms to help computers to make these choices seems feasible.

This leads to the most vicious phrasing of the Frame Problem, which I term the *Significance Problem*: How can a model formalize only the important things? The problem is not how to formalize everything, but how to formalize only things that make a difference in the operation of our story rewriting tutor. Standards of significance must be made explicit for a model to use them, yet this is often a fairly subjective judgment that is open to interpretation. This is a call for participatory design: *humans* can recognize what is significant. Teachers definitely have training and rationale in rating plot quality. Yet, a crucial component of rating plots consists of unconscious skill and fine interpretation rather than a list of rules that are easily formalized. Practically, StoryStation could have multiple judges read possible interpretations of the story, rank the events by the importance, and try to find some measure of reliability for this. Realistically, we would find it doubtful that such a method would actually be reliable and it would require considerable human effort. A much more natural task for teachers it to give a story a single holistic ranking for plot.

4.3 Teacher-Designed Rating

The Frame Problem is endemic to all formal models and so can never be completely overcome, yet a participatory model of development can ground a formal system effectively in the practice of informal humans. Traditional artificial intelligence techniques sometimes fail since they tried to find or invent an explicit methodology. We have two distinct advantages which make our project more tractable. First, full story understanding is not required for automated plot analysis. The analysis must simply be able to rate a student's story using some metric of plot recall and comprehension. Once this rating is done, feedback on how to improve the story needs to be given to the student. Second, instead of having to invent a methodology for plot rating, we can use participatory design principles to have a teacher create the plot rating metric[16].

A Scottish schoolteacher with a history of interest in StoryStation and forty years experience with the Scottish National Curriculum, decided to help us design the plot rating metric. The Scottish schools have no rigorous guidelines for the rating of story plot for the story rewriting task. After some reflection, the teacher formulated a plot rating scheme that corresponded to one she used in class. She felt that no precise algorithm could be written, but that much of the task of rating a story's plot comes from a complex play of factors carefully judged implicitly by a teacher. However, while she could not give us a set of formal rules, she did produce a set of informal guidelines about what aspects of a story merit in a particular rating that could be used by other raters. She volunteered to rate half of rewritten stories in the corpus, and other raters independently rated all the stories, using her ratings as a base-line to see if the story rating metric was reliable. Given her insights into the inherently informal nature of plot rating, with her agreement it was decided that one feasible approach would be to create a statistical model of how the raters rated the stories by using a machine learner. A *machine learner* is a computational system that, given a set of correctly classified data (called "training data") and features of that data, creates a statistical model of the classification scheme that it can then automatically apply to new data, which is called the "test data." In this way, the informal nature of story grading by teachers can be effectively modeled without recourse to frames or explicit rules, instead focusing on modeling how a teacher assesses the plot of a story as they often do in a classroom setting.

5. PLOT RATING METRIC

The rewritten stories were rated for plot by three different raters. The second author (Rater B), who is also a storyteller, and a non-expert (Rater C), the first author, rated all of the stories. The teacher (Rater A) who designed the metric rated half the stories, as this was all her time allowed. The following scale, as dictated verbatim by teacher, was given to all raters to use as their guidelines for rating stories:

1. *Excellent*: An excellent story shows that the child understands the "point" of the story and should demonstrate some deep understanding of the plot. The student should be able to retrieve all the important links and, not all the details, but the right details.

2. *Good*: A good story shows that the student was listening to the story, and has recall of the main events and links in the plot. However, the student shows no deeper understanding of the plot, which can often be

Class	Probability	Number of Class
1 (Excellent)	0.175	18
2 (Good)	0.320	33
3 (Fair)	0.184	19
4 (Poor)	0.320	33

Table 1: Distribution of Plot Ratings in Corpus

detected by the writer leaving out an important link or emphasizing the wrong details.

3. *Fair*: A fair story shows that the child is missing more than one link or chunk of the story, and not only lacks an understanding of the "point" but also lacks recall of vital parts of the story. The fair story does not really flow.

4. *Poor*: A poor story has definite problems with recall of events, and is missing substantial amount of the plot. Characters will be misidentified and events confused. Often the child writes on the wrong subject or starts off reciting only the beginning of the story.

5.1 Validity

Between Rater A and Rater B there was a Cronbach's α statistic of .8840 and a Kendall's τ_b statistic of .821. Between Rater B and C there was a Cronbach's α statistic of .9327 and Kendall's τ_b statistic of .869. These statistics show our rating scheme to be fairly reliable. Since Rater B rated all the stories and was not involved the implementation of the plot analysis software, her ratings were used as the gold standard. The distribution of these ratings are shown in *Table 1*. It was felt that using Rater A, who designed the metric, would be infeasible since she only rated half the stories. She also designed the metric itself, and we were interested in how valid the metric was across independent raters who were given only the explicit text of the metric. Since Rater C implemented the plot analysis algorithm and designed the details of its statistical and linguistic operations, it would have been unfair for his ratings to be used as gold standard to test his algorithms.

It appears that there is not much agreement between Rater A and Rater B, with only 39% (18 of 46) agreement. Upon closer inspection, it is clear that Rater A was just systematically more harsh than Rater B. Stories that Rater B would classify as fair would be classified by Rater A as poor. In fact, almost half (13 of 28) of their disagreements fall into this category. Upon further inspection of the stories that were the sources of disagreement, it seems that Rater A would never give partial credit to an incomplete story and would always mark it as poor, even if it was very near completion. Rater B often gave it partial credit, marking incomplete stories as fair. Rater A in general tended to grade down, often grading a rating scale one less than that of Rater B. However, large digressions between the two were rare, as an excellent story was never identified as a poor story and only twice were stories ranked as good by Rater B marked as poor by Rater A. There were only 3 two-rank differences in total. The rest of the errors were evenly dispersed between disagreements over excellent and good stories (7 disagreements) and good and fair stories (5 disagreements). The final results are easily explainable, if one accepts the explanation that more than four-fifths of their disagreements was just Rater A marking one rank lower than Rater A. When asked about this, Rater A felt that Rater B's ratings were equally valid, just one degree more lenient than her own, and that she had graded incomplete stories as poor by default but could have graded them more partially.

Rater C and Rater B, who rated all the stories, had a high agreement of 77 percent (79 out of 102). Rater C was consistently marking stories higher than Rater B, with 19 stories marked higher and only 4 lower than Rater B. Rater C tended to be less harsh than Rater B, although he was more in agreement with her than Rater A. The largest difference is that stories rated as fair by Rater B, Rater C would tend to mark as good. He also tended to rank as fair those that Rater B would mark as poor. Only in very rare circumstances (once) did Rater C rate a story fair that Rater B would rate as poor. Again, the raters would usually differed by one ranking, and in one direction. The largest area of disagreement was in between good and fair stories, with 11 disagreements, although disagreements between fair and poor stories came in a close second with 8 disagreements. Fair and excellent stories had only 3 disagreements. Overall, Rater C apparently was often in close agreement with Rater B, just giving more partial credit.

Overall, the rating scheme is far from perfect, but fairly reliable. It is imperfect enough for human raters to have relatively large disagreements on the rating scale, yet good enough that the raters tend to not disagree by more than one ranking. Some of the differences between the raters could be explained by their experience as well as style. The least experienced rater rated stories leniently, while the most experienced rater rated the most harshly.

5.2 Comparison to Psychology

It is interesting to note that very little of the work into understanding children's narratives has taken any of the relevant psychological and discourse literature of children's narratives into account. In the tradition of Charniak[5], children's stories are either viewed as the same as adult stories or as simple adult stories. As emphasized by the teachers, the varied and complex linguistic development a child is expected to go through that differentiates their story writing and story understanding ability from that of adults. Since we need to deal with surface text and the characteristics of children's linguistic development around the age of 10-12 (the age of most users of StoryStation), these differences need to be taken into account. Children *develop* the ability to use narrative over time. One well-known model of narrative development in psychology is the Applebee model[2]. Although this psychological model exists, we preferred to use participant design to get a valid experienced teacher's perspective. However, comparing that model to the teacher's for consistency would be useful. The typical levels of a child's ability to understand and rewrite plot are as follows in Applebee's developmental model[2, 14]:

1. *heaps*: A series of unrelated referents and events. This shows the basic structure of stories to be "bare" referents and events, even if they are unconnected. This parallels the *Poor* category in the teacher-designed metric.

2. *sequences*: A series of events linked by a single referent, usually with some type of similarity relation between events. This shows the development of a *Fair* or *Poor*

story into a *Good* story.

3. *focused chains*: The focus now follows the main character. Note that if the events are recalled correctly (as in a *Good* story), the plot follows the main character.

4. *narratives*: Expansion of the focus to other elements of the story in an orderly fashion, as well as elaboration on themes. This understanding of the theme seems to reiterated in the teacher's description of the understanding of a point in being crucial to *Excellent* stories.

By the average age student using StoryStation should have the capability to fully use narratives. Still, these developmental levels are useful as some students may have a slower development of narrative use. These levels show that the teacher's implicit knowledge of stories reflects well the findings of developmental narrative psychology. The collected corpus reflected these levels of development, with stories ranging from unrelated events and characters to an understanding of the causal flow of events and point of a story. It should be noted that the rating metric used by the teacher is preferable to simply using the developmental model, since the rating metric is designed in the social context of the use of StoryStation and specifically with the story rewriting task in mind. There are differences between the psychological model and the teacher-designed rating metric, with the most obvious one being the focus of the teacher on not just recall and character following but getting the crucial "point" of the story.

6. THE DESIGN OF THE AGENT

Decisions have to be made about what levels our computational model of plot analysis operates on and what level the teacher needs. It is possible to analyze the stories on a number of levels. First, the similarities of the word distributions between stories can be compared. A good rewritten story would presumably at least share many of the same words as the original story. However, using this as the only measure of plot quality is problematic. The teacher-designed metric places high priority on actually recalling events in the correct order. Our system needs to extract and analyze events from the raw text of the rewritten stories and compare them for temporal order. For example, a story that had all the events described backwards in time (so that the beginning would be at the conclusion) would not be detected a statistical analysis of word context. The use of synonyms would violate word context. It was found perfectly acceptable by the teacher for the main character Nils to be called a "boy," yet any statistical analysis of those words would not realize that relation and count the use of the word "boy" as a wrong-headed variance in the rewritten story. The event extractor must also have enough built in flexibility to recognize this. Our system should at least attempt to conflate synonyms to get a proper grasp of statistical word similarity. The idea of a "point" is very difficult to formalize, although it may be possible for it to be an emergent factor from proper event recall and ordering combined with the use of certain key words.

6.1 Events

Instead of describing the different elements of the story ("Introduction," "The Climax," and so on) in the tradition of Propp, we consider the plot elements to have only two different categories, events and entities, so that they can be automatically extracted from text[22]. Entities are nouns that include animate characters such as "boy" and "geese" and inanimate objects such as "coins" and "cities." Events are composed of the interactions of entities, such as the "boy throwing the coin," which is composed of a "boy" and a "coin" connected by "throwing." Events are predicates, and entities are arguments to these predicates, with the predicate being named by an verb indicator. For example, "the boy throws the coin" maps to *throws(boy,coin)*. Together with an ordering over time, these can form a non-quantified events such as *throws(t=13,boy,coin)*, where t is a integer variable denoting the order of the event in the child's story. A sentence may map onto one, multiple, or no events, such that the sentence "Nils walked down the beach while the stork slept" would map to *walk(t=1,Nils,beach), slept(t=2,stork)*. Two stories are said to match if they are composed of the same ordering of events.

6.2 Extracting and Comparing Events

The event calculus was extracted from raw text of children's storing by layering natural language processing components using an XML-based pipeline. For this particular XML pipeline we used the LT TTT (Language Technology Text Tokenization Toolkit) and LT XML[11, 28]. A full description of the pipeline is beyond the scope of this paper, see [12, 13] for details. The guiding constraints used was that the event extractor had to operate over the often ungrammatical raw text of children and the process had to be fully automatic. First, words were tokenized, sentences separated, words tagged for their part-of-speech, and then a rule-based anaphora resolver was used to resolve pronouns to common or proper nouns. Events were then extracted, using the Cass Chunker to extract tuples of verbs and nouns in sentence chunks[1]. The original story, a transcript of story as told to the children, was used to create an event series for the original story. A plot comparison algorithm steps through the events of the original story and compares each of them to the rewritten story. The events are compared to each other by checking to see if the events are equal first in number of entities, and then in time. If the entities of the event are not equal, WordNet[8] is used to automatically extract a synonym set for the entities being checked, and membership is checked. If the temporal order is different the entities are also checked. Assuming that the automatic event extraction could have left out crucial information, the event entities are also checked to see if they are present in the raw text of the story. This checking process leads each rewritten story to be characterized as a series of events, with each event having a number of binary attributes as iterated above. After analysis of the results of comparing events in differing ways[12], the binary attributes found to be most useful were "Entity present in raw text of rewritten story" and "Synonym used to describe entity in rewritten story." The more tight constraints of temporal order and precise event-by-event comparison were found only to hurt performance[12].

6.3 Statistical Analysis of Rewritten Stories

We used *Latent Semantic Analysis*[18] (LSA) scores as our metric of word distribution similarity. LSA provides an approximation of "semantic" similarity based on the hy-

pothesis that the semantics of a word can be deduced from its context in an entire document[10]. LSA compares the words of a document to the words of another document, and produces a score ranging between 0 and 1.0 as a similarity score between one story and another story. Before the advent of LSA, this was often taken as the measure of word co-occurrence in two stories. Since this would penalize synonym usage among other things, in LSA before the comparisons are taken the stories are projected unto a "semantic space" representative of the English language. The space is created by reducing a large corpus of texts to a smaller subspace, so that the subspace has a smaller dimensionality than the original space. The texts used in our experiment are the required reading of 12th-grade students from the USA as collected by TASA[18]. Once both stories are projected to this reduced subspace, a cosine comparison takes place to measure their similarity. In the reduced subspace, similar words such as "coin" and "money" are collapsed into one dimension, and so not penalized by the similarity score.

In comparison to the event extraction methods, LSA keeps closer to the actual data in the students text since it is made from probabilities in the original data itself and the TASA reading level corpus with a dimensionality of 200[18]. There are severe limitations to the use of LSA. First, it does not take word order into account, so the sentence "Nils jumped on a goose" and "The goose jumped on Nils" would have a similarity score of 1.0. As Hickmann suggests, the very flow of narration in children's discourse in based on interplays between regularities such as grammar and semantic roles[14]. The presence or absence of specific events and temporal order are not captured adequately using only word frequency distributions over the entire document, regardless of how complex the mathematical processing (such as sub-space reduction in LSA) upon these documents is. Despite these shortcomings, similarity scores from LSA are useful since they provide information about the words used in the text that are stripped out in event extraction.

6.4 Combining the Methodologies

Every story can now be represented by our plot analysis agent as a series of events (an event structure) with binary attributes and a discretely normalized LSA similarity score. One principled way to model a teacher's plot rating ability it to let a machine-learning algorithm learn what features are good predictors of a teacher's grade. The machine "learns" by inspecting a portion of corpus of graded rewritten stories (the training data), counting the occurrences of each feature to build a statistical model of each rating class. It first inspects the rating of the story and inspects the event structure's attributes and LSA similarity score. If the values of this event are presence in a particular rewritten story with a particular rating, it adds these to its model of that rating. Once each story has been iterated through in this manner, a statistical model of each rating class is built. In a simple hypothetical example, the machine-learner could learn that 60% of good stories have LSA score greater than .50 and the presence of the event $find(Nils, coin)$. It could then notice that 70 percent of all poor stories had a LSA score of less than .40 and were missing more than half of the events. When given new stories to rate (the test data), the machine learner inspects its event structure and LSA similarity score to determine, given its statistical models of each rating class, what the most probable rating the story belongs to. This type of machine learner is called the *Naive Bayes* machine learner, since it is based on Bayes Theorem and the "naive" Conditional Independence assumption. The mathematical details of the model are available[12], and a full analysis of various other machine-learning techniques in this task is available[13].

Class	Precision	Recall
Excellent	0	0
Good	0.433	0.879
Fair	0.455	0.263
Poor	0.917	0.667

Table 2: Automated Plot Analysis: Precision and Recall per Rating

7. RESULTS

Using the Naive Bayes machine-learner and ten-fold cross validation, the automated plot analysis program correctly identified 51.46% of all the stories in the 103 story corpus. Ten-fold cross validation is when the 90% of the corpus as training data and 10% as test data, and iterating this division through the whole corpus ten times, so every story is used a test data exactly once.

While at first this score may not seem impressive, in light that for the human raters the agreement ranged between 39-77%, it is a moderate success. In fact, had the automated plot analysis agreed completely with Rater B, the results would most likely have been overlearned instead of an accurate assessment of the objective validity of the rating model due the amount of human disagreement in using the rating scheme. What was found in observing human teacher's ratings was that variations in using the rating scheme tended to be systematic, and the same should be expected to hold in with the automated plot analysis. A table of the precision and recall scores is presented in *Table 2*.

Upon closer inspection, the machine-learner rated the poor stories with a fair degree of accuracy (66.7%) and good stories with a high degree of accuracy (87.9%) stories. However, it experienced a severe amount of difficulty processing both excellent and fair stories, consisting identifying less than 20% of each rating. On a more fine-grained level, the automated plot analysis separated the stories into two large clusters, with 67 stories being identified as good stories (29 correctly) and the poor (24) stories. Only 11 stories were identified (5 correctly) by the analysis as fair and only 1 story was identified (and incorrectly) as excellent. Almost all excellent stories were misidentified (17 out of 18) as good stories, with the last one being misidentified as fair. The fair stories were also mostly identified as good stories, with 13 fair stories misidentified as good stories, and only 1 fair story identified as a poor story. A few good stories were misidentified as either excellent (1), fair (2) or poor (1), but for the most part (87.9%) they were identified correctly. Poor stories were mostly identified correctly (66.7%), although a surprisingly amount (8) were identified as good stories and (3) fair stories.

8. DISCUSSION

The plot analysis system works, albeit in a limited fashion. Its ability to separate good stories from poor stories is

remarkable, and would come in useful to a teacher. While its behavior is too limited to allow it to automatically grade rewritten stories, the real issue is of adapting the ratings given by the automated plot analysis system to the context of StoryStation. Its ability to accurately identify poor stories allows it to identify the pupils that are in need of help. Using the automatically extracted event structure, the agent could remind the pupil that they may have forgotten a part of the plot or give a pupil who is struggling to write the story suggestions about what part of the plot to focus on next by using its internal event structure for the rewritten story. The agent could suggest to pupils who are having difficulty to ask for help from a teacher. The teacher thought that this ability to find students who are struggling with the plot of the story would the system's greatest boon to teachers, since often those students who are struggling often are the last to ask for help. The ability to separate good from poor stories relieves much additional tedium from the teacher, since having to manually distinguish those stories is a time-consuming task. Instead of having to wait for a teacher's comment, the plot analysis agent allows the student to recall the plot more effectively and so concentrate on developing other facets of writing.

The behavior of the automated plot analysis system gives important clues about the limits of our current model, and the entire approach of modeling story plots. It is not surprising that poor stories can be identified with relative ease by computers. As noted by the teacher-designed rating scale, *poor* stories tend to be much shorter than longer stories and have differing word distributions, aspects taken into account by using LSA similarity scores. They are more likely to be missing large chunks of events, a phenomena easily recognized by our automatic event extraction. The main characteristic of good stories is that they posses "recall of the main events and links in the plot." This would be easily identified by the automatic event extractor. However, the automatic plot analysis system fails to discover both fair and excellent stories. Part of the reason for the poor performance is that both types of stories were less frequent in our corpus that good and poor stories. Another reason may be that in that in the rating scheme the teacher identified both excellent and fair as relying on notions of "understanding the point" of the story. The concept of a "point" is obvious to humans such as teachers and to students (once they "get" the point), but difficult if not impossible to formalize in a computational system. The few attempts that have been made by researchers like Wilensky [29] rely upon using AI planning techniques that can not handle the idiosyncratic logic of children's stories. As regards excellent stories, it is not surprising that the story system should fail so, as excellent stories involve not just retrieving "all the details", but the "right details" as noted by the rating scheme. Our event extractor by nature extracts "all the details," and apparently its attempts to discover the "right details," most likely due to the small size of the corpus. Excellent stories often use creative words which are not caught by LSA similarity scores of WordNet in the algorithm. Fair stories, by virtue of missing events, are easily detected by the event extraction, but may share many of the same words and events as good stories, thus confounding the system. The idea of "flow" is also difficult to formalize. Overall, our system succeeds and finding and extracting events, but can not tell if a student got "the point" of the story.

9. CONCLUSION AND FUTURE WORK

As Charniak discovered, appearances can be deceiving: children's stories are not easy for a computer to understand[5]. There are a number of crucial errors in this viewpoint. First, children's stories are *complex*. Children are not simplified versions of adults, children are living human beings that are dynamically developing, and their stories embody complexity as much as any other natural phenomena. Their pronoun usage may be simpler, but their syntax tends to have an ungrammatical but endearing complexity all of its own that foils traditional parsers. There is still much work to be done, but such a plot-analysis agent in StoryStation is considered by the teachers we have consulted with to be a social good and can serve as a useful demonstration of artificial intelligence and natural language processing to the "real world" teaching domain. By not dealing with artificial frame representations and modeling a human teacher, our system is able to both bypass to an extent the Frame Problem and its attendant knowledge engineering bottleneck.

One should not forget there are social aspects to allowing a computer to grade stories. Do we really want our children to be graded by machines that may not recognize important factors in plot such as creativity? Given the current ratio of teachers to students, it makes sense that a teacher would like some automation to alert them to students that are having serious problems with their writing, so that the teacher can concentrate on the struggling student. It also allows the teacher to focus their attentions on teaching the human element of writing, emphasizing creativity and quality writing technique. Although machines may never be able to grade stories with the same care as humans, one can imagine interactions between children and machines that allow both to grow and reach new heights of complexity and skill in development. StoryStation is not intended to grade stories, but allow the students access to an artificially intelligent agent that can produce constructive feedback on the stories to the pupil. With our plot analysis system, an animated agent may remind a child of a forgotten point or help a child with a particularly difficult word. Due to the statistical nature of our system, new stories (either automatically rated within a reasonable measure of correctness or human rated) can be used to improve the system. Thus, the child's writing will in turn influence the behavior of the agent, allowing the agent to learn how human writing works better. However, before such work takes place, the main future work of this project is a longitudinal evaluation of the educational effectiveness of StoryStation in the field. The software has been installed in a primary and a secondary school where it will be used by classes of 10-11 year old and 12-13 year old pupils as part of their classroom writing instruction over the course of two school terms. Qualitative interview data is being collected in order to characterize the pupils' and teachers' experiences with the software in terms of motivation and their perception of the effect it has on learning. Quantitative analysis of log file data and pre-test and post-test writing samples will also be carried out.

The search for a perfect model of human stories on computers may be doomed to failure by its nature, but the study of the interaction and mutual development of humans and computers will lead to more insightful research in artificial intelligence and participatory design. Children's stories led artificial intelligence researchers to realize that they were in the dark about essential aspects of intelligence. The ability

to narrate events, to tell stories, is fundamental to human intelligence. While our plot analysis system here is far from perfect, its use of computational linguistics and teacher participation allowed us to make progress on what appears to be an impossible task. We hope this type of design will serve as a framework for future work into narration and computation. Researchers have barely scratched the surface the depths of human story understanding. Yet, a practical and participant-centered approach to this problem can lead to socially useful application like StoryStation.

10. ACKNOWLEDGMENTS

Special thanks to Senga Munro for her assistance in designing the plot rating scheme, rating the stories, and for her advice. Also this project would not have been possible without the pupils of Methilhill and Torbain Primary Schools in Kirkcaldy, Scotland. Lastly, thanks to all the teachers and students who have been involved in the development of the StoryStation project.

11. REFERENCES

[1] S. Abney. Chunks and dependencies: Bringing processing evidence to bear on syntax. In J. Cole, G. Green, and J. Morgan, editors, *Computational Linguistics and the Foundations of Linguistic Theory*, pages 145–164. 1995.

[2] A. Applebee. *The child's concept of story.* University of Chicago Press, Chicago, 1978.

[3] B. Britton and S. Glynn. *Computer Writing Environments: Theory, Research, and Design.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1989.

[4] J. Burstein, D. Marcu, and K. Knight. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, pages 32–39, 2003.

[5] E. Charniak. *Toward a Model of Children's Story Comprehension.* PhD thesis, MIT, 1972.

[6] D. Christian and M. Young. Comparing Cognitive and Computational Models of Narrative Structure. Technical report, North Carolina State University, 2001.

[7] A. Druin. Co-operative inquiry: Developing new technologies for children with children. In *Proceedings of CHI'99.* ACM Press, 1999.

[8] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[9] L. Flower. *The construction of negotiated meaning: A social cognitive theory of writing.* Southern Illinois Univeresity Press, Carbondale, IL, 1994.

[10] P. Foltz. Latent Semantic Analysis for Text-Based Research. *Behavior Research Methods, Instruments, and Computers*, 1996.

[11] C. Grover, C. Matheson, A. Mikheev, and M. Moens. LT TTT - A Flexible Tokenisation Tool. In *Proceedings of the Second Language Resources and Evaluation Conference*, 2000.

[12] H. Halpin. The Plots of Children and Machines: Statistical and Symbolic Analysis of Narrative Texts. Master's thesis, Edinburgh, October 2003.

[13] H. Halpin. Automatic Analysis of Plot for Story Rewriting. In preparation. Available upon request at http://www.semanticstories.org, 2004.

[14] M. Hickmann. *Children's Discourse: person, space and time across language.* Cambridge University Press, Cambridge, UK, 2003.

[15] M. Kyng. *Users and computers: A contextual approach to design of computer artifacts.* PhD thesis, Aarhus, 1995.

[16] M. Kyng. *Computers and Design in Context.* MIT Press, Cambridge, MA, 1997.

[17] S. Lagerloff. *The Wonderful Adventures of Nils.* Doubleday, Page, and Company, Garden City, New York, 1907.

[18] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 1997.

[19] C. Levi-Strauss. *Structural Anthropology.* Basic Books, New York, 1963.

[20] J. McCarthy and P. Hayes. Some philosophical problems from the standpoint of Artificial Intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 4. 1969.

[21] M. Minsky. A framework for representing knowledge. In P. H. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, New York, 1975.

[22] V. Propp. *Morphology of the Folktale (translated by Laurence Scott).* University of Austin of Texas, 1968.

[23] J. Roberston and P. Wiemar-Hastings. Feedback on children's stories via multiple interface agents. In *International Conference on Intelligent Tutoring Systems*, Biarritz, France, 2002.

[24] J. Robertson. Experiences of child-centred design in the StoryStation project. In *In Proceedings of Interaction, Design, and Children*, Eindhoven, 2002.

[25] J. Robertson and B. Cross. Children's perceptions about writing with their teacher and the StoryStation learning environment. *Narrative and Interactive Learning Environments: Special Issue of International Journal of Continuing Engineering Education and Life-long Learning*, 2003.

[26] D. Rumelhart. Notes on a schema for stories. In D. Bobrow and A. Collins, editors, *Representation and Understanding: Studies in Cognitive Science.* Academic Press, New York, 1975.

[27] M. Scaife and Y. Rogers. Kids as informants: Telling us what we didn't know or confirming what we knew already? In A. Druin, editor, *The Design of Children's Technology: How we Design, What we Design, and Why.* Morgan Kaufman, 1999.

[28] H. Thompson, R. Tobin, D. McKelvie, and C. Brew. LT XML: Software API and toolkit for XML processing, 1996.

[29] R. Wilensky. Points: A theory for the structure of stories in memory. In W. Lehnert and M. Ringle, editors, *Strategies for Natural Language Processing*, pages 345–375, Hillsdale, N.J., 1982. Lawrence Erlbaum.