Towards Automated Story Analysis Using Participatory Design

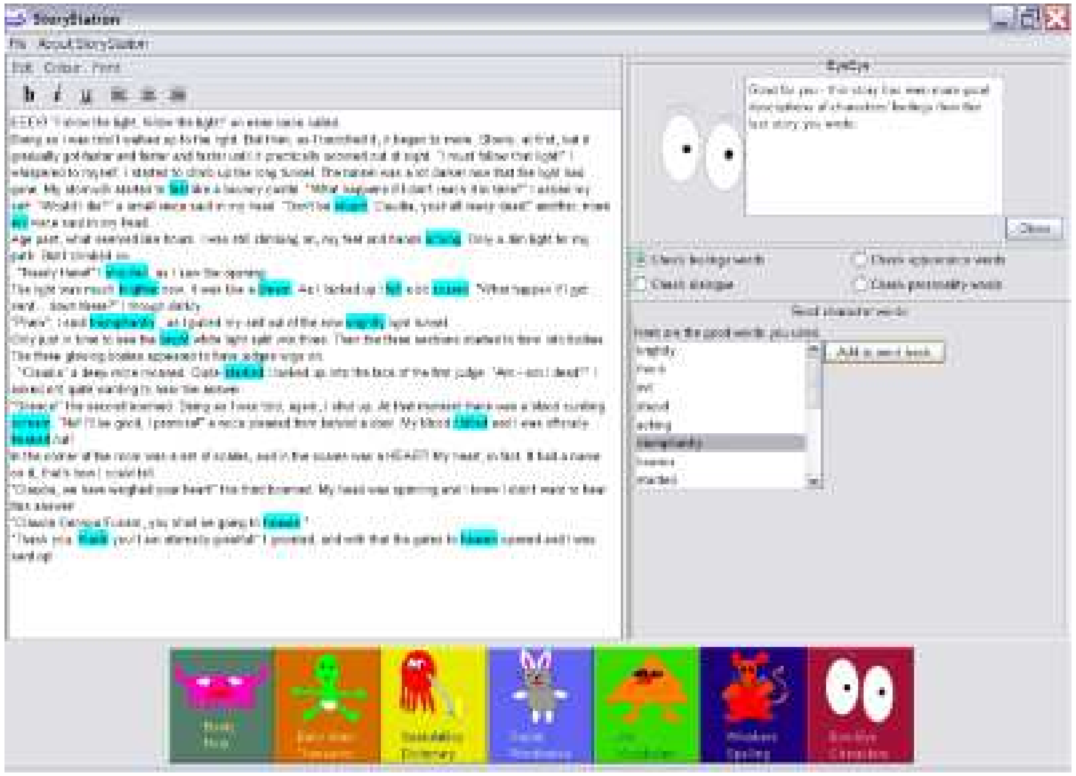
Harry Halpin, Johanna D. Moore, and Judy Robertson

HCRC/ICCS University of Edinburgh

What is StoryStation?

- An intelligent tutoring system to aid children (age 10-12) in their writing.
- Used participatory design with students and teachers.
- Story writing viewed as a set of skills, each skill (such as word choice, spelling, plot evaluation) is given an animated agent.
- Recent results (2001, Qualifications and Curriculum Authority) have shown children have trouble properly ordering the events of a story and having satisfying resolutions.
- Being test-driven in Scottish schools by the Edinburgh-Stanford Link.

Pinky the Plot Analyzer



Sample Output

JK Rowling, Harry Potter and the Philosopher's Stone, Ch. 5

Well done! The descriptions of your characters are very good. Have a look at your sentences and notice where you can join two sentences with a good joining word like ”but” or ”because”.

I think your story would be even better if you wrote more about how the characters feel.

I think your story would be even better if you included more dialogue.

What do you think the characters would say to each other?

Here are some of the good words you used: quivered, wriggled, twisted, cauldrons, tottering, parchment, scarlet, swarthy, beware.

The Problem with Plot

Traditionally story understanding has been pursued via coding of “common-sense” knowledge into machines and then use inference and planning mechanisms with to “understand” a story.

This is difficult with fiction, with *talking* storks and *magic* coins.

Knowledge Engineering Bottleneck: To manually code all common-sense knowledge into a database is a time-consuming specialist skill, even for domains such as children’s stories.

Significance Problem: The modeler does not know the unintended consequences of leaving something out of their model.

Story Rewriting Task and Plot Critiquing

- **Story Rewriting**: Students are read a story (the *exemplar story*) and rewrite it (the *rewritten story*), allowing them to focus on skills such as word choice and syntax without the cognitive load of plot creation.

- **Plot Analysis**: Compare an exemplar story to a rewritten story to assess quality and provide feedback to student. What was the quality of the rewritten story? Did the pupil remember the events? Similarity is not the whole story.

- **Story Classification**: Need to classify stories on general plot quality, but need also be able to *interpret* and use the results of plot classification.

- Not used for automated grading, instead enables agent to provide help to student or to encourage student to ask for help from the teacher.

Automatic and Statistical Approach

Requirements: No knowledge engineering, fully automatic acquisition of plot structure from text of children's stories.

Solution: Use tools from computational linguistics to extract a semantic representation automatically from rewritten story.

Model Human Judgment: Use a corpus of stories graded by experts to create a statistical model of plot quality for each story using a machine-learner (currently Naive Bayes). Count how often the presence or absence of each part of the semantic representation is correlated with a particular level of plot quality.

Mix Statistical Components with Symbolic Components: Use a word-frequency based approach since extracted representation may leave something important out.

Corpus Creation

Size: 103 rewritten stories from one exemplar story from two classes.

Rating Scale: Teacher devised 4-tier ranking scheme based on ability of the student to recall events and "understand the point."

Excellent	Understands "point" and recalls "right details."
Good	Recalls events but no "deeper understanding of plot."
Fair	Lacks more than one "chunk" of the story.
Poor	Lacks "substantial" part of plot.

Similar to Applebee model of narrative development, although teacher more focused on "understanding the point."

Raters

- Stories graded by 3 independent raters. One chosen as gold standard for machine-learning.
- Absolute agreement of raters low (Avg. raw agreement among raters 58%).
- Yet disagreement was **highly** systematic.
- Overall raters agree when one story is better than another, but may differ on absolute rating given.
- The two raters with the highest disagreement (39%) had Cronbach's α of .86.

Overview of Plot Analysis

For each rewritten story:

- Automatically identify events using tools from computational linguistics.
- Use Plot Comparison Algorithm to compare events with exemplar story's events.
- Perform LSA (Latent Semantic Analysis) comparison with exemplar story text.
- Use results of above as features in machine-learner to rate stories automatically.

Event Calculus

Motivation: Need to automatically identify "events" in students' stories.

Uses a simplified version of the **event calculus** (temporally ordered non-quantified predicate calculus) since it captures relevant aspects of plot, such as *temporal order, characters,* and *events*.

Automatically converts text to an event calculus format via a **pipeline** that includes POS tagging, pronoun resolution (rule-based), and chunking.

Verbs of each chunk converted into event (predicate), nouns into entities (arguments). Works on **ungrammatical** text.

Automatically keep track of time using algorithm that keeps track of tense (Allen, 1984). Most children pick a tense and stick with it (Hickmann, 2003).

Example

Sentence	Event
Nils lives in Sweden	*live(t=1, Nils, Sweden)*
and he is always playing with geese on his mountain.	*play(t=1,Nils, geese, mountain)*
He sees a bird,	*see(t=2, Nils, bird)*
so he walks down big mountains.	*walk(t=2, Nils, mountain)*
What a mystery!	*Nothing produced*
It was a stork.	*is(t=3, stork)*

Plot Comparison Algorithm

Motivation: Need a measure of difference between the events in the rewritten story and exemplar story.

The algorithm iterates through the exemplar story's events looking for matches in each event of the rewritten story.

Events are matched by matching their components one at a time, starting with the event name and then matching entities.

Uses WordNet to find out if synonyms were used and uses a "now-point" to determine if events have been temporally inverted.

Example

Algorithm Result: (E=exact match,S=synonym,N=no match, T=In order, N=Not in order)

Exemplar	Rewritten	Result
throw(t=1,Nils, coin)	toss(t=3,coin)	S, N, E, T
see(t=2,Nils, city)	appear(t=4,city)	N, N, E, T
enter(t=3,Nils, city)	walk(t=4,Nils)	N, E, N, T
ask(t=4,Nils, merchant)	say(t=5,merchant)	S, N, E, T
say(t=4, Nils)	say(t=6,boy)	E, S, T
leave(t=5,Nils)	go(t=8,Nils)	$E, , T$
disappear(t=6,city)	vanish(t=7,city)	S, E, N

Produces a feature vector of these results for each rewritten story.

Classifying Stories

Machine-learning allows the significance of the presence or absence of particular events to be found. Features include similarity discovered with word frequency comparison in a reduced subspace (LSA) and the results of the Plot Comparison Algorithm.

Here's current best results using 10-fold cross-validation (90% training, 10% test data) with only a few representative results:

Machine Learner	**% Correct**	**Cronbach's α**
Naive Bayes	**54.37%**	.78
Avg. Rater Agreement	58.37%	.88

Results

This confusion matrix shows the distribution of ratings output by *Naive Bayes*.

Machine-Learner

Human

Classification	1	2	3	4
1 (Excellent)	0	**17**	1	0
2 (Good)	1	**29**	2	1
3 (Fair)	0	13	5	1
4 (Poor)	0	8	3	**22**

Naive Bayes collapses all categories except “good” and “poor,” but has a low error spread for these ratings. “Fair” and “excellent” stories require “getting the point”.

Discussion: Future Work

The ratings need more precise specification.

Need a larger corpus, test domain independence better.

More complex temporal order extraction and better temporal ontology, especially for phenomena such as inter-character conversation .

Exactly how do we generate agent feedback on plot?

How to capture “getting the point” of a story?

Can this be generalized to make a computationally tractable model of plot quality for story understanding?

Discussion

Teachers find classification useful: identify missing events, story quality, identify temporal inversions.

Use of events allows agent to give feedback to student about what events are missing, characters misidentified, and such.

Difficult to tell what level of abstraction is correct for a given task and corpus, but preliminary results are encouraging for this task.

The algorithms for plot extraction and plot comparison are completely automated, no knowledge engineering required.

Automated plot analysis of the story rewriting task solves a real-world problem by mixing knowledge representation, participatory design, and statistics.