Assessing MultiPlot Stories: from Formative Analysis to Computational Metrics

Pablo Gervás, Eugenio Concepción, Gonzalo Méndez
Facultad de Informática
Universidad Complutense de Madrid
Madrid, 28040 Spain
{pgervas,econcepc,gmendez}@ucm.es

Abstract

Recent interest in story generators capable of combining more than on plot line into an elaborate story have been handicapped by the lack of either theoretical material or quantitative metrics to ascertain the quality of outputs of such attempts. The present short paper postulates a set of metrics designed to capture some of the insights elaborated during a formative evaluation of an existing attempt at plot weaving.

Introduction

The mechanics of how to combine more than on plot line into a rich story have become a subject of interest in storytelling research in recent times. Solutions have been proposed to address the task (Fay 2014; Porteous, Charles, and Cavazza 2016; Gervás 2014a; 2018). Yet there is a shortage of either theoretical material or quantitative metrics to ascertain the quality of outputs of such attempts. The present short paper postulates a set of metrics designed to capture some of the insights elaborated during a formative evaluation of an existing attempt at plot weaving. The metrics are calibrated against qualitative evaluations by human judges and tested over outcomes of baseline solutions for plotline weaving. The metrics emulate the observations made by human judges in that they consider separate sets of positive and negative features. The metrics are designed to identify features that at some point in the formative evaluation have been deemed by some human judge to either add or detract to the perceived value of a story. The overall judgment on a given story must be extrapolated from the corresponding collection of features.

Related Work

The three topics considered relevant for this short paper are prior solutions for plot line combination, quantitative metrics for stories and formative evaluation of plot weaving.

Plot Line Combination

The systems reviewed here all combine a number of "plot lines" in some form, but each uses a different terminology for referring to them. To facilitate description, we consider an abstract concept of *plot line* as a sequence of plot elements, each describing an event relevant to the structure of the story, with the possible addition of a set of roles that characters play in the event.

Fay (2014) considers a plot weaving algorithm that builds new stories combining plot lines for a set of given character types. The system finds the character models best matching the given types, retrieves narrative threads associated in the corpus with those models, and finds the best combination of those narrative threads into a single story, ensuring that characters' plots are compatible and that the resulting timeline is consistent.

Porteous, Charles, and Cavazza (2016) presents an interactive storytelling system that constructs stories with multiple interleaved plot lines. Their system constructs the stories dynamically using a plan-based approach in response to set of input parameters that drive the number of plot lines to be interleaved and the relative time spent on presentation of each subplot.

The StoryFire (Gervás 2018) system generates stories inspired by the movements of pieces in a chess game. This system combines concepts of narrative thread – sequence of predicates affecting a given piece – and plot line – a linear sequence of abstract labels for plot-relevant events that may describe an interesting story line. In this case, the plot line (usually single) is used to inform an interweaving of narrative threads for different characters.

Computational Metrics for Stories

Existing previous work on quantitative metrics for stories has not addressed multi-plotline stories explicitly. The work in (Gervás 2014b) describes a number of metrics to quantify a set of desired structural features over narrative renderings of game logs, and it focuses on issues such as coverage of the game, and features like redundancy and continuity of the composed discourse. Earlier works focused on metrics for story novelty (Peinado et al. 2010) and related concepts such as similarity between stories (Hervás et al. 2015). In particular, (Hervás et al. 2015) describes a calibration process based on comparing results on the metric against human judgement.

Formative Evaluation of Plot Weaving

The work of (Concepción, Gervás, and Méndez 2020) explores baseline solutions for weaving together a set of plot

Proceedings of the 12th International Conference on Computational Creativity (ICCC '21)
ISBN: 978-989-54160-3-5

templates into stories where scenes from the different templates appear interleaved. A plot template would correspond to the plot line we are considering – these plot templates include additional information on roles played by the characters. Several procedures for combining plot templates are described, some based on existing literary techniques (Communicating Vessels, Chinese Boxes) and some presented as baselines for computational approaches to the task (subplot concatenation, subplot alternation, and random mixing). A formative qualitative evaluation of 10 story examples is included. This evaluation includes qualitative analyses by human judges of the stories in question, where specific features that add or detract to the perceived value of the story are discussed.

Automated Emulation of Human Assessment of Plotline Weaving

The present short paper describes a set of metrics designed to capture in a numerical form the insights that arose from the formative evaluation presented in (Concepción, Gervás, and Méndez 2020). This formative evaluation uncovered insights at two different levels: features perceivable in stories that are considered valuable by human evaluators, and types of knowledge about the story that are being brought into play by human evaluators when making such judgments.

Insights on Desireable Features in Multi-Plotline Stories

The insights that have been considered relevant for the quality of plot line weaving, and susceptible of numerical formulation are described below.

The comments in the formative evaluation made it clear that there are two features of the stories that play an important role in the perception that human judges have of the quality of their weaving: the valence of characters (whether they are good or evil) and the level of activity conveyed by each scene.

Evaluators praised stories where sub-plots have been combined merging villain with villain or hero with hero.

They also praised stories where descriptive scenes from one plot line were interleaved with descriptive scenes from another, and active scenes were interleaved together. This intuitively leads to a story that switches from a more descriptive mode to a more narrative mode at one point, and the subplots that make it up allign in that sense.

Another feature that was considered relevant is the rhythm of alternation between sub-plots when they are interwoven. Evaluators praised stories in which the rhythm of alternation between subplots – how many scenes from each subplot are told together before switching to the other – matches the perceived impression of activity for the story. If scenes are active, and significant events are happening in each sub-plot, switching between sub-plots can happen every few scenes; whereas if scenes are descriptive and nothing much actually happens in each one, more time should be spent on each sub-plot before switching to another.

Two further features were mentioned as positive for some stories: the existence of an overarching plot for the story that starts and ends the story, and the appearance of a complete sub-plot as an insertion within another.

It is important to note that, when asked to assess stories, human judges did not resort to scoring them or ranking them, rather made a set of observations on each story. These observations were either positive (identifying positive features in the story) or negative (identifying negative features in the story). The metrics that we are proposing follow this same pattern.

Knowledge about Stories Relevant to Multi-Plotline Assessment

The analysis of the formative evaluation suggested that valence of the characters and level of activity of scenes are relevant features that need to be made available to a system hoping to assess multi-plotline stories. Therefore the existing set of resources was hand annotated with values for these features. A baseline annotation was carried out over the templates for sub-plots as a first approximation. In this way, the relevant information is tied in to each plot template.

Valence for characters in a given scene was annotated with a value of -1 for characters performing evil actions and 1 for characters performing good actions. A valence value of 0 is assigned by default to all other characters.

Level of activity of scenes was annotated by adding a flag to scenes in a template that involved some relevant action. The rest of the scenes are considered descriptive.

Quantitative Metrics for Multi-Plotline Weaving

The system as it stands can parse stories written from text files in a particular format into a representation in terms of templates built of scenes. It also allows construction of new stories by combining a number of plot lines using the baseline computational strategies described in (Concepción, Gervás, and Méndez 2020). In both cases the representation that is obtained allows for the automated compilation of numerical data for character valence and activity based on the annotations described.

The procedure constructs four different types of vectors of numerical values for each story:

- *vectors of character valences*: for each character, compile the sequence of valence values for the scenes in the story
- *vectors of sub-plot alternation*: for each span of the story corresponding to a different sub-plot, note which template it comes from
- *vectors of alternation rhythm*: for each span of the story corresponding to a different sub-plot, note its length in number of scenes
- *vectors of matching scene activity*: for each of the spans in the alternation rhythm sequence, compile the count of active scenes

Over these vectors, a number of features considered by the human evaluators can be computed automatically. In all cases, the philosophy is to identify features that at some point in the formative evaluation have been deemed by some human judge to either add or detract to the perceived value of a story.

[CL1] West is a scientist. Lily is West's wife. West lives with his family in the countryside. [DO1] West is a wealthy man; Mary loves West; West loves Mary. [CL2] West dreams with creating artificially a boy. [DO2] Benson is the new steward of West Manor. [DO3] Mary is West's sister. Mary falls in love with Benson. [**DO4**] Benson seduces Mary. [CL3] West creates artificially a boy. West names the boy as Nemo. [CL4] West teaches Nemo to behave as an ordinary boy. [**CL5**] Nemo accidentally kills Mary. Nemo discovers that he is not human. Nemo leaves home. [**DO5**] Benson is discovered by Mary while stealing money. [**CL6**] Nemo steals food in a farm. [CL7] West looks for Nemo. [DO6] West fires Benson. Mary is happy again. [**CL8**] West finds Nemo in an abandoned house in the town. West deactivates Nemo. West is sad and sorry.

Table 1: Story 3 combines a Creation of Life (CL) subplot with a Destructive Outsider (DO) subplot. Scene labels from each subplot are shown in [square brackets], in **bold** if negative in valence. Active events underlined.

[DO1] Augustine is a innkeeper. Claude is the major of the town. [DO2] Lupin is an foreigner that arrives to the town. Lupin hires a room in Augustine's inn. [CL1] Hubert is a magician. Hubert wants to create humans from animals. [CL2] Hubert captures a wolf and transforms it into a human. Hubert names the new creature as Lupin. [CL3] Hubert teaches Lupin how to behave like a human. [CL4] Lupin becomes a werewolf during the night. [**CL4**] Lupin goes to the forest. Lupin kills a deer. [CL5] Hubert follows Lupin's trail. [**CL6**] Hubert discovers Lupin near a farm. Lupin kills Hubert. [DO3] Lupin is very kind with Augustine. [**DO4**] Lupin becomes a werewolf during the night. Lupin kills Augustine. [**DO5**] Lupin tells Claude that a wolf attacked Augustine. [BO1] Claude is sick. Jack arrives in town. [BO2] Jack tells Claude that the town is in danger. Claude does not trust Jack. [BO3] Jack reveals that there is a cursed stone in the town. The curse is making people get sick. [BO4] Jack destroys the stone. Jack dies. Jack saves the town. [BO5] Claude becomes healthy. [DO6] Pearce and Alain tracks the forest. [DO7] Lupin becomes a werewolf during the night. [**DO8**] Lupin attacks Pearce and Alain. Alain shoots Lupin. Lupin dies. [DO9] Pearce and Alain become heroes in the town.

Table 2: Story 9 combines a Creation of Life (CL) subplot with a Destructive Outsider (DO) subplot.

The automatic identification of the following features has been implemented:

- overarching plot (vector of sub-plot alternation starts and ends with the same sub-plot)
- inserted sub-plot (sub-plot appears only once in vector of sub-plot alternation)
- spans with regular interweaving rhythm (a given value of alternation rhythm is maintained over a number of transitions between sub-plots)
- rhythm matched to activity (either slow rhythm for spans with low activity, or high rhythm for spans with high activity)

In addition, the values for valence of characters are used to build an overall pattern of alternation between valences is built for a story. This allows the establishment of distinctions between stories that end events with negative valence (tragedies) and stories that end in events with positive valence (comedies, rags to riches stories, overcoming the monster stories. . .).

Discussion

The proposed metrics are calibrated against the inspiring stories and tested over automatically generated stories.

Calibration over Inspiring Stories

The results for the proposed metrics over the inspiring stories considered in the formative evaluation of (Concepción, Gervás, and Méndez 2020) are presented in Table 3.

The application of the metrics to these stories is intended as a calibration exercise, to test whether the metrics indeed capture the intuitions that inspired them. Observations on story quality are not considered because the formative evaluation used as reference did not explicitly consider them.

The set of stories includes examples of accepted strategies used in literary text (Chinese Boxes inserts a complete subplot as a single span within another, Communicating Vessels interleaves several subplots with different rhythms). These strategies represent instances of complex weaving strategies that are considered valuable. The metrics clearly identify the Chinese Boxes strategy in stories 6, 7, 9 and 10 (by design the results include both overarching plot and inserted plots).

Story 9 has a span of identified rhythm (*rhytSp* = 1) has a similar situation towards its end (two contiguous spans of 4 scenes) and these also happen to include no activity so they are recognised as a slow pace segment (*slow* = 1) of the story, with relatively slow subplot alternation matching scenes low in activity. Story 3 has a similar situation with spans of 3 scenes, but the activity in that case is not regular. Examples of these stories are shown in Tables 1 (Story 3) and 2 (Story 9). The examples have been chosen using the same subplots to allow comparison of the differences in structure between the resulting complete story.

The Communicating Vessels strategy exercises greater freedom in the way it combines subplots, allowing it to choose whether to include an overarching plot (story 2) or not (stories 5 and 8). Because it interweaves subplots more freely, it can result in a higher number of regular rhythm spans (*rhytSp*, see story 8).

The Alternation strategy by design imposes a fixed rhythm of alternation (*rhytSp*) leading to a single span of regular rhythm of the same size as the story (*spSiz*, see Story 1).

The Random strategy has the potential to replicate the freedom of the Communicating Vessels strategy, as shown by the similar values shown by the metrics for *rhytSp* and *spSiz*.

The patterns for valences show a marked tendency towards positive endings (7/10) over negative ones. This is a natural consequence of the nature of the templates considered (only 1/4 ends on a negative valence). Overall there is a marked tendency to start stories on a negative note (the classical solution of starting with a conflict to be resolved). This again is a result of the set of templates used.

Testing over Generated Stories

The results of testing the proposed metrics over a larger set of automatically generated stories are shown in Table 4.

StID	*Strt*	*#pl*	*ovar*	*ins*	*rhytSp*	*spSiz*	*slow*	*fast*	*valences*
6	B	2	✓	✓	0	[]	0	0	[-1, 1, -1, 1]
7	B	3	✓	✓	0	[]	0	0	[-1, 1, -1]
9	B	3	✓	✓	1	[2]	1	0	[-1, 1, -1, 1]
10	B	3	✓	✓	1	[2]	0	0	[-1, 1, -1, 1, -1, 1]
2	V	2	✓	✗	1	[2]	0	0	[-1, 1]
5	V	2	✗	✗	0	[]	0	1	[-1]
8	V	3	✗	✗	1	[3]	0	0	[-1, 1]
1	A	2	✗	✗	1	[12]	0	0	[-1, 1]
3	R	2	✗	✗	2	[2, 2]	0	0	[-1]
4	R	2	✓	✗	1	[3]	0	1	[-1, 1, -1, 1]

Table 3: Results for metrics for inspiring stories in the formative evaluation. Stories are grouped by strategy: *StID* id in (Concepción, Gervás, and Méndez 2020), *Strt* is strategy used – A alternating, R random, V Communicating Vessels, B Chinese Boxes –, *pl* is number plots, *ovar* overarching plot, *ins* inserted plot, *rhytSp* spans with regular rhythm, *spSiz* sizes of regular rhythm spans, *slow* slow pace span, *fast* fast pace span and *valences* valence pattern.

Strt	*#pl*	*ovar*	*ins*	*rhytSp*	*slow*	*fast*	*v-s*	*v+e*
C	3	0	0	12 (2)	12	0	14	11
C	2	0	0	9 (2)	9	0	11	7
A	3	2	0	8 (19)	0	0	5	10
				6 (18)				
				4 (17)				
A	2	5	0	1 (14)	0	0	9	9
				8 (13)				
				6 (12)				
				5 (11)				
R	3	5	1	1 (10)	7	7	10	6
				1 (7)				
				4 (5)				
				6 (3)				
				14 (2)				
R	2	12	3	2 (6)	2	1	11	10
				2 (5)				
				2 (4)				
				6 (3)				
				6 (2)				

Table 4: Results for the metrics over a set of automatically generated stories: *Strt* is strategy used – C concatenation, A alternating, R random – *#pl* is number of plots, *ovar* number of overarching plots, *ins* number of inserted plots, *rhytSp* number of spans with regular rhythm – number of spans (size of the span) – , *slow* number of stories with slow pace spans, *fast* number of stories with fast pace spans, *v-s* number of negative valence story starts and *v+e* number of positive valence story ends. All values over 20 runs for each case.

The stories are generated using the three baseline computational strategies described in (Concepción, Gervás, and Méndez 2020): concatenation, alternation and random. Results are reported as totals over a set of 20 generated stories for each strategy.

The values for the metrics serve to highlight the shortcomings inherent in the baseline weaving strategies. The Concatenation strategy allows neither overarching plots nor inserted plots. Regularities in rhythm arise by serendipity whenever (at least two of) the subplots involved have the same length. Because the spans involved are always long, the pace of switching between subplots is identified as slow. The Alternation strategy allows overarching plots in certain cases (50% of the time when two plots are used, namely when one of the is longer than the other) and does not allow inserted plots. Due to its nature, it generates a single span with a regular rhythm of alternation of the same size as the story – which varies depending on the templates employed. The templates available do not include continuous sequences of active scenes, so identifying patterns of activity at that rhythm is almost impossible. The Random strategy does allow both overarching plots and inserted plots, and it allows the appearance of spans of regular rhythm in different patterns. The results of this strategy sometimes include several spans of interweaving at different rhythms. In this sense, it is the only of the computational strategies tested that can emulate the behaviour of the Communicating Vessels reference strategy. The Random strategy does have shortcomings of its own in that it is altogether blind to any features that it might be introducing.

With respect to valences, the reported outputs are built using a larger set of templates, with higher prevalence of evil acts towards the end (in 4 out of the 7 templates used). This leads to a higher average of evil ends (slightly over 4 out of 10 as opposed to the 3 in 10 of the hand-crafted stories). The set of templates used is chosen at random, which leads to a more even distribution (5 out of 10 in average) of positive vs. negative beginnings. The values for the hand-crafted stories may have been affected by the original decision to rely on a restricted set of similar templates throughout to make it easier to perceive changes in structure resulting from different weaving strategies.

An example of a generated story is shown in Table 5. This story presents a number of the features that are identified by the metrics proposed in this paper. There is a regular rhythm span with three scenes from the Split Personality Comic subplot (SP1 to SP3) followed by three from the Creation of Life (CL3 to CL5) subplot. There is another regular rhythm span with two scenes from the Split Personality Comic subplot (SP4 and SP5) followed by three from the Creation of Life (CL6 and CL7) subplot. There is no overarching plot, because the story starts with a scene from the Creation of Life subplot and ends with a scene from the Split Personality

[CL1] Scott dreams of bringing to life his recently deceased wife Julia. [CL2] Scott uses a spell to bring to life the portrait of Julia. [SP1] Edward is a physicist researching parallel universes. [SP2] Edward accidentally switches with himself in another universe. In the parallel universe he is Hans, married Martha and working in a farm in the countryside. [SP3] Edward does not know how to take care of the animals and causes various messes. Edward falls in love with Martha. [CL3] Scott talks every evening with Julia. Scott tells Julia that Scott loves Julia. [CL4] Scott falls in love with Edward. Edward falls in love with Scott. Scott tells Julia Scott loves Edward. [**CL5**] Edward visits Scott. Julia is jealous of Edward. Julia takes Edward and disappears. [SP4] Martha realises that Edward is not himself. Edward confesses to Martha what has happened. Martha asks for the original Hans to return. [SP5] Edward begins working on a system to replicate his experiment. Edward manages to communicate with Hans in his original universe. Edward and Hans work together to fix the problem. [CL6] Scott searches for Julia across the castle. [**CL7**] Scott discovers Julia near a farm. Scott kills Julia. [SP6] Edward and Hans manage to reopen communication between universes. Edward returns to his universe and leaves to find his own Martha.

Table 5: Generated Story 2 for Random combination of 2 subplots: combines a Creation of Life subplot and a Split Personality Comic subplot.

Comic subplot. This, together with the choice of subplots, implies that the story has a positive start and a positive end.

The story also includes a number of additional features that are not covered by the metrics but which are clearly relevant for the task. The formative analysis of the stories in (Concepción, Gervás, and Méndez 2020) identified the problem of inconsistency between the life spans of characters unified between two subplots – characters that die in the final story as required by one of the subplots but then continue active as required by another. The story in Table 5 presents two instances of similar phenomena: (1) Edward falls in love with Martha (SP3) and then Edward falls in love with Scott (CL4), and (2) Edward kidnapped (CL5) but remains active without having been released (SP3 and SP4). This type of issue needs to be addressed in further work. It will very likely require further enrichment of the resources with information on when characters are restricted in movement or fall in love.

It is also interesting to note, that, given the peculiarities of Split Personality Comic subplot the characters Edward and Hans are both the same person and separate characters. This complicates computation of this type of consistency restrictions.

Intended Application of the Metrics

The metrics reported here are proposed as a first step towards devising a set of informed weaving strategies that aim to produce stories that exhibit the features identified as desirable. This short paper reports on the enrichment of the underlying resources and the development of the metrics and constitutes a preliminary result. Further work will explore weaving strategies that may take advantage of both of the contributions reported here (enriched knowledge resources and computational metrics for desired features) to achieve multi-plotline stories exhibiting the features deemed valuable by the human judges during the formative evaluation used as inspiration.

Solutions for the automated extraction of the knowledge resources from corpora of stories will also be explored.

Conclusions

This short paper reports these preliminary results on the metrics. The proposed metrics built automatically do serve to identify the features in the inspiring set of stories that they were intended to capture. The baseline solutions for plot weaving considered produce unimpressive output scored low by the metrics. Small peaks in the score do seem to match serendipitous good features observable in the output stories.

Ongoing efforts exist to develop plot weaving solutions to optimise these metrics. It is hoped that such plot weaving solutions will lead to significant improvements in the outcomes.

Acknowledgments

This paper has been partially funded by the projects CANTOR: Automated Composition of Personal Narratives as an aid for Occupational Therapy based on Reminescence, Grant. No. PID2019-108927RB-I00 (Spanish Ministry of Science and Innovation) and InVITAR-IA: Infraestructuras para la Visibilización, Integración y Transferencia de Aplicaciones y Resultados de Inteligencia Artificial, UCM Grant. No. FEI-EU-17-23.

References

Concepción, E.; Gervás, P.; and Méndez, G. 2020. Exploring baselines for combining full plots into multiple-plot stories. *New Generation Computing* 1–41.

Fay, M. P. 2014. *Driving story generation with learnable character models*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Gervás, P. 2014a. Composing narrative discourse for stories of many characters: a case study over a chess game. *Literary and Linguistic Computing* 29(4).

Gervás, P. 2014b. Metrics for desired structural features for narrative renderings of game logs. *Journal of Entertainment Computing*.

Gervás, P. 2018. Storifying Observed Events: Could I Dress This Up as a Story? In *5th AISB Symposium on Computational Creativity*. University of Liverpool, UK: AISB.

Hervás, R.; Sánchez-Ruiz, A. A.; Gervás, P.; and León, C. 2015. Calibrating a metric for similarity of stories against human judgement. In *ICCBR (Workshops)*, 136–145.

Peinado, F.; Francisco, V.; Hervás, R.; and Gervás, P. 2010. Assessing the novelty of computer-generated narratives using empirical metrics. *MINDS AND MACHINES* 20(4):588.

Porteous, J.; Charles, F.; and Cavazza, M. 2016. Plan-based narrative generation with coordinated subplots. In *European Conference on Artificial Intelligence (ECAI 2016)*, volume 285, 846–854. IOS Press.