Summaries Can Frame—But No Effect on Creativity

Leonid Berov
Institute of Cognitive Science
University of Osnabrück
49076 Osnabrück, Germany
lberov@uos.de

Abstract

Framing, the accompanying information that sometimes comes provided with a work of art in order to explain the artist's intention, the creation process or to present the artwork in a special light, has been theorized to improve the appearance of creativity of the generating act (Charnley, Pease, and Colton 2012). One type of highly regarded framing in the literary domain is the *critique*, an interpretative work which provides a functional, thematic and/or symptomatic condensation of the essence of the primary text, basically, a summary.
The present paper empirically tests whether computationally generated narratives can, too, be framed through functional summaries, and whether this framing indeed contributes to the system's perceived creativity. To do so, it employs the functional unit (FU) summary approach conceived—but never fully implemented—by Lehnert (1981), in order to summarize a story generated by a storytelling algorithm. It compares the performance of FU summary with other approaches, and based on this data evaluates whether better summaries can also serve as better framings, as well as whether better framings increase a system's perceived creativity. Our results indicate that (1) FU based summary performs around human level, (2) better summaries are indeed judged to be better framings, but that (3) neither of these two factors have a significant effect on perceived creativity. Based on this we conclude that further scrutiny and empirical study is required to understand how framing can be harnessed for computational creativity.

Introduction

Charnley, Pease, and Colton (2012) describe how artists often present their work in a special light; an endeavour that seems to contribute to the artwork's quality and the creators perceived creativity. An iconic example is Marcel Duchamp's infamous piece 'Fountain': A ready-made urinal that was submitted (unaltered but for the signature 'R. Mutt') under this title to an avant-garde exhibition in 1917. Creativity can not be attributed to the creation of the object—it is not even the product of the artist's own work. Nor can it be attributed solely to the refined aesthetic sensibility required in spotting its appeal, since similar pieces of plumbing do not seem to have garnered comparable fame. Indeed, the artwork itself was never shown during the exhibition, yet sparked an important artistic debate about its rejection. In an editorial on the case, dadaist Louise Norton wrote:"Whether Mr Mutt with his own hands made the fountain or not has no importance. He CHOSE it. He took an ordinary article of life, placed it so that its useful significance disappeared *under the new title and point of view—created a new thought* for that object" (Norton 1917, italics mine for emphasis). This sentiment makes clear that the creative act, here, rests mostly in the uncommon viewpoint and interpretation that the artist had provided.

Providing a work with accompanying information about itself, its purpose or creation has been called *framing*, and is one of four crucial types of *generative acts* that a creative system might perform (Colton, Charnley, and Pease 2011). Charnley, Pease, and Colton (2012) suggest that creative systems would benefit from performing framing-type generative acts: "As with human artworks, the appeal of computer creativity will be enhanced by the presence of framing". Following the working definition of our field this means that "unbiased observers would deem [such a system] to be [more] creative" (Colton 2012). This understanding seems to be plausible from an analytical perspective: it holds for many cases where human creativity is concerned. Yet, to the best of our knowledge, it has never been tested in a generative context: When a computational system creates an artefact, is it indeed perceived as more creative, when it adds a decent framing to the package? The present paper makes a first attempt to address this question empirically in the storytelling domain.

Framing in the Story Domain

So far, three types of framing have been distinguished in the literature: (1) *Motivation* i.e. what lead to the creation of an artwork, (2) *intention* i.e. what is the purpose/foreseen effect of the artwork and (3) *process* i.e. how was the artwork created (Charnley, Pease, and Colton 2012). These can be seen as related to the four factors of creativity identified by the '4P' model (Rhodes 1961). The first two capture the influence of *person* (individual factors) and *place* (societal factors), while the third one relates noteworthy details about the *process*. The missing factor is *product*, and we suggest that it be included as a framing type in its own right. Product type framing can be a re-description/re-interpretation like in our incipient example, but also just an accentuation of spe-

cific properties of the artwork.

In the case of narratives, condensed product-type information is often understood as a summary (or in more elaborate cases a critique). Summarization necessarily implies an abstraction from the subordinate and the particular, therefore the possible types of summary also depend on the level at which abstraction is performed. Based on the meta-narratological reflections in Eder (2010), three product-intrinsic levels can be differentiated[1]: (1) *Artefactual* i.e. descriptions concerned with the narrative's structure and form, (2) *representational* i.e. descriptions concerned with the narrated content and (3) *thematic* i.e. interpretations concerned with higher meaning like symbolism and messages. The high cultural regard for literary criticism, which is essentially the discipline of providing other people's narratives with thematic framing[2], shows the potential for this line of thought. Based on these observations, the present paper empirically investigates whether summaries can be used to frame narratives.

Functional Summarization

Previous work demonstrated the utility of an approach called Functional Unit Analysis (details see below) for both summary generation and aesthetic evaluation (Wilke and Berov 2018) in a computational storytelling system. While the general feasibility was demonstrated on a case study, it remained unclear how the quality of the resulting summary compares to human level and the state of the art. As its final contribution, the present paper performs a quantitative evaluation of the FU summarization technique.

Summarizing, our argument structure will thus be the following:

1. For one story, we create summaries using different approaches (including Functional Unit Analysis) and evaluate their quality comparatively.
2. Based on this comparison we investigate whether better summaries can serve as better framing for their story.
3. Departing from this analysis we determine whether a better framing can enhance a computational system's perceived creativity stronger than a worse framing.

Related Work

Two main tasks are distinguished in text summarization: *extractive summarization* aims at extracting the main information-bearing sentences in the source text, while *abstractive summarization* generates text not contained in the input, based on some sort of reasoning about the content (Gambhir and Gupta 2017). If the summary is supposed to be used as a framing, too, abstractive summarization seems a more promising route since it has the potential of introducing new content instead of just reordering already known text.

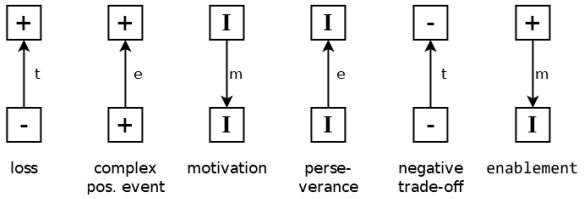

Figure 1: A sample of primitive units adopted from (Lehnert 1981). I denotes an intention, $+$ and $-$ are vertices of positive and negative affect.

Data-driven Summarization

Presently, the common approach to natural language generation tasks like abstractive summarization is the use of deep sequence-to-sequence neural networks based on an LSTM encoder-decoder architecture, which are trained on large corpora of text in a supervised way (Sutskever, Vinyals, and Le 2014). The current state of the art was achieved by See, Liu, and Manning (2017) by extending a vanilla LSTM approach with mechanisms for copying words from the source by a technique called pointing, and considering the coverage of the already summarized input during generation.

It should be noted, that work performed in this context is concerned with the analysis of argumentative, and not narrative, text. This is an important distinction, since the former is much more likely to carry its 'point' on the textual surface (like e.g. in the incipient sentence of a news article) than in a deep structure, like in the case of narratives (as e.g. a fable's moral). Especially in data-intensive approaches, which extrapolate summarization rules based on the text—that is a corpus of existing summaries—these differences between genre can be expected to lead to problems in generalization to the present use case. This applies to the work of See at al., too.

Functional Unit Summarization

A very different approach is the functional unit (FU) model, which was proposed as a tool for the abstractive, analytical summarization of narratives by Lehnert (1981). It operates on a graph representation of plot and works by identifying strategically significant portions of the plot called complex FU, which are expected to be points of high relevance for summaries. Lehnert's plot graphs can contain three different types of vertices, which represent mental states resulting from characters' perceptions of the events of the plot. The mental states that can be contained in a vertex are positive (denoted: $+$) or negative ($-$) affect, or intentions (I, with neutral affect). Positive states describe any event which is appraised with positive emotions by a character, while negative states describe the inverse. Intentions are courses of

[1] While Eder's analysis focuses on character, we see no reason why it should not be applicable to the larger narrative context. The forth level he proposes, dubbed symptomatic, is excluded here because it is product-external and focuses on descriptions better captured at the place factor of the 4P model.

[2] With this **we do by no means intend to demean the critique** as a highly valuable analytical and interpretative text type, and a creative endeavour in its own right. **Rather, we want to elevate framing** which in its best instantiations might aspire to be a little more like a critique and less like a plain, referential summary.

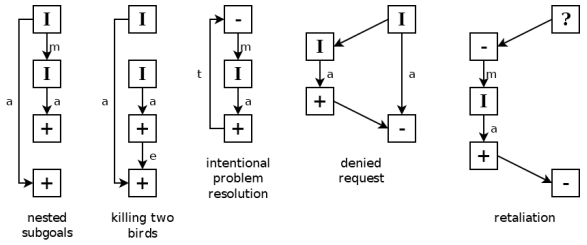

Figure 2: Examples of complex FUs adopted from (Lehnert 1981). '?' represents wildcard vertices.

actions the character has committed to as a reaction to their perception. Vertices are interconnected by edges that describe how these states are related to each other. They can be of the following types: motivation, actualization, termination, equivalence or inter-character edges. Based on this formalism, Lehnert defines "15 legal pairwise configurations" (primitive FUs) that act as an alphabet: they capture semantically meaningful two-state configurations like e.g. 'motivation' or 'loss' (see Fig. 1). From these primitives an open set of complex units can be constructed, which capture more intricate plot configurations, e.g. 'denied request' or 'retaliation' (see Fig. 2). New complex FU can be easily constructed, however, Lehnert does not provide a formal definition of which situations should count as complex FU, and which not. For the present purposes we make do with the FU already introduced in Lehnert (1981).

A story is analysed by transforming the story-text into the introduced graph-representation, and then detecting all FUs contained in it. When this is done, a *connectivity graph* is built by using the different instances of FUs as vertices, and connecting them with edges wherever two unit instances share one or more vertices in the plot graph. To generate a summary, the units contained in the connectivity graph get translated into natural language by using template-like *generational frames* which are supplied to the program for each unit type. Into the frames, information about the specific instance of a unit is fed, allowing the frames to generate text including the characters involved in the unit or some other unit-specific context (for an example see Fig. 3).

An attempt at implementing this procedure for the analysis of human-made stories yielded modest results (Goyal, Riloff, and others 2013), due to the complexity involved in translating literary text into the proposed graph representation. The natural language processing required for this includes complex interpretative tasks like event-based discretization, intention and emotion detection as well as the identification of causality relations—problems not commonly addressed in research. Fortunately, this impediment can be avoided when dealing with computer generated plots. If the information required for the graph are created by an algorithm in the first place, then no natural language interpretation is required to extract them, and the only task left is the generation of an appropriate graph. The feasibility of computationally modeling enough narrative phenomena to be able to create most of Lehnert's primitive unit alphabet has been demonstrated recently (Wilke and Berov 2018, see

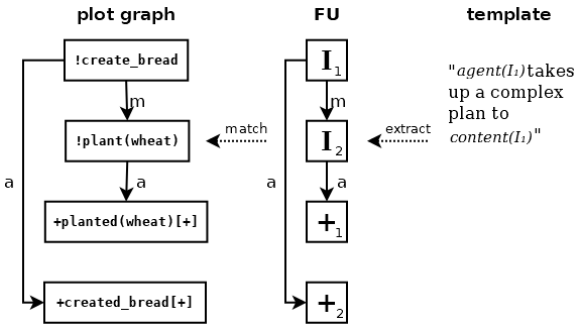

Figure 3: An extract from a plot graph, the FU 'nested subgoal' matching it, and the FU's generational frame which, when applied, will generate the text "hen takes up a complex plan to create_bread".

here for more technical details).

Study Design

Our study has three interconnected goals: (1) evaluate how well FU Analysis-based summarization performs, (2) establish whether a better summary provides a better framing for a story, and (3) test whether a better framing leads to higher creativity ratings for the generating system. In such a setting conclusions can be drawn only in the case that significant differences are present in the compared summaries in the first place. For this reason we saw fit to employ three different approaches to generate these summaries. The technology under test was FU Analysis-based summarization (condition F). A lower bound was expected to be established by employing data-driven abstractive summarization as these approaches are not specialised in narrative text, which should lead to a sub par performance (condition D). An upper bound can be established with safety by generating the summary through a human subject, as human-level performance has so far not been computationally surpassed (condition H).

In order to address the three goals above, a questionnaire with three sets of questions: about summary quality, framing quality and perceived creativity needed to be established. Human subjects could then be presented with $(story, summary_x)$-pairs[3], with $x \in \{F, D, H\}$, and asked to answer the questionnaire. To reduce the number of required participants a within-subject design was chosen, where each subject successively observes and rates all three conditions. This is beneficial, because it strongly reduces noise due to intersubjective differences. The order of presentation of conditions was counterbalanced to prevent interaction effects like primacy, habituation or simply boredom. A comparable setup has been demonstrated to perform well in previous work (Berov and Kühnberger 2018).

[3]Subjects were always first presented with the story and then the summary, on the same page. Future experiments might see fit to control for this order.

Experimental Conditions

The fairy tale "The Little Red Hen"[4] was used as target story because its re-implementation in a storytelling system was demonstrated to be a suitable basis for comparable empirical evaluation (Berov and Kühnberger 2018). The benefit of recreating an existing story is that a high-quality textual surface form already exists, which can be used as input for data-driven computational summarization techniques, in lieu of the poor prose generatable with off-the-shelf NLG systems.

We chose the system presented by See, Liu, and Manning (2017) to generate condition D, because it presents a recent state of the art in abstractive summarization and provides both, code as well as a pretrained model, online[5]. To the best of our knowledge, no data driven work has been dedicated to computationally summarize stories, and no accepted corpora exist for this domain, which would have allowed us straightforwardly training a specialized model. For this reason, we saw fit to employ the model pretrained by the authors on news text. The length of the thus generated summary can not be independently controlled since it is one of the features learned by the model[6]. For this reason the length of the data-driven summary was used to determine the target length for the two other conditions: 50 ± 5 words.

Condition H was generated by asking a human subject with higher education to carefully read the fairy tale and write a summary of the required length. The subject was given no further information about the experiment or its constraints, and was not provided with example summaries deemed felicitous by us. We want to explicate that this convenient route is grounded in the assumption that any such individual can be expected to have extensive experience with text summarization, and replicating the performance of even the worst human sample could count as success in a computational system.

Condition F was generated by recreating the plot of TLRH using our simulative storytelling system (Berov 2017) and automatically applying FU summarization as described above, with the text of the FU templates slightly adopted in order to fit the target word count.

Since the summaries are ultimately intended to be used as framing, the required minimal linguistic changes were performed on all three conditions in order to turn them into first-person explanations, i.e. by prepending the clause: "I wanted to write a story about". The resulting explanations read:

- **Condition D**: I wanted to write a story about a little red hen that lived on a farm with a dog, a pig and a cow. The dog, the pig, and the cow said they were too tired to help. When the bread was done, she put it in the oven to bake.
- **Condition H**: I wanted to write a story about a hen that lived on a farm with 3 animal friends. She worked in the garden, while the animals did nothing but sleep. After much work growing wheat and making bread, the hen told her friends she would eat the bread alone, since nobody had helped her.
- **Condition F**: I wanted to write a story in which the hen takes up a complex plan to create bread, the pig, the cow and the dog deny the hen's request for help and the hen retaliates against the pig, the cow and the dog by punishing them.

As predicted above the machine-learning based summary (condition D) is of low quality; in particular it demonstrates a lack of understanding of the finer mechanics of the bakery trade and, in our opinion, fails to capture the story's main points. This is felicitous since keeping a low-quality exemplar allows the validation of the employed questionnaire by checking its sensitivity for low quality and establishing a lower-bond comparison point for the condition under test.

Survey Questionnaire

A questionnaire has been created in order to estimate the perceived creativity C of the storytelling system, the suitability of a text as the framing F of a story, and the quality of a text as summary S (see Fig. 4). Each item is a statement to which participants have to indicate their agreement using a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

The items C1 through C3 assess the creativity of the system by inviting feedback on its product; focussing on quality, typicality and novelty, which are the criteria brought forward by Ritchie (2007). Items C4 through C6 assess the creativity of the system by inviting feedback on the process, focusing on perceived imagination, appreciation and skillfulness, which are the criteria brought forward by Colton (2008).

In accordance with the discussion of Charnley, Pease, and Colton (2012) in our section "Framing in the Story Domain", items F1 through F4 elicit assessment on a text's capability to (1) put an artefact in a specific light, (2) enhance an artefact's aesthetic value, (3) provide a plausible intention for the artefact and (4) frame the creative process.

Surprisingly, the qualitative evaluation of summaries seems not to be extensively theorized, so that items S1 through S4 were created in a more ad-hoc manner. They were designed to elicit assessment on a text's capability to (1) representationally capture content, (2) provide condensation through abstraction, (3) still achieve good coverage, and (4) distill the thematic dimension of the text.

The items are combined in three thematic groups, and in each condition the groups were presented to the subjects in a randomized order to balance any potential interaction effects.

After all the items of each condition, participants were also presented with an optional free text field allowing them to answer the following question: "If the explanation affected how you perceive the story, please explain in a few words in what sense it did so", aimed at eliciting qualitative data to understand why certain summaries are better suited as framing than others.

[4] www.home.uos.de/leberov/tlrh.htm

[5] https://github.com/abisee/pointer-generator

[6] Abigail See, personal communication

```
C1  The text was an interesting story.
C2  The text was a good exemplar of a fairy tale story.
C3  The text was novel.
C4  The algorithm that created this story seems to have imagination.
C5  The algorithm that created this story seems to engage in an aesthetic evaluation of its own work.
C6  The algorithm that created this story seems to be skillful in story-writing.

F1  The explanation changed the way I see the story.
F2  The explanation enhances the story's value as a work of art.
F3  The explanation provides the story with a meaning or a message.
F4  The explanation reveals some of the reasoning that went into the creation of the story.

S1  The explanation summarizes the content of the story well.
S2  The explanation conveys an abstract understanding of the story.
S3  The explanation captures the main points of the story.
S4  The explanation explicates the meaning or message of the story.
```

Figure 4: Questionnaire used to evaluate all three conditions, presentation order of the three groups was randomized for each participant.

	Summary	**Framing**	**Creativity**
D	1.79 ± 0.70^{1}	1.69 ± 0.61^{1}	3.00 ± 0.66^{1}
F	3.86 ± 0.85^{2}	3.22 ± 0.92^{2}	3.18 ± 0.67^{1}
H	3.74 ± 0.82^{2}	2.49 ± 0.86^{3}	3.14 ± 0.66^{1}

Table 1: Survey results: perceived quality of text as summary and framing, and perceived creativity of generating system (mean $\pm$ std) for the three conditions. Superscripts indicate groups with statistically significant differences (at least at the $P \leq 0.0001$ level).

Results

An online survey platform was used to carry out the study. 36 participants were recruited from the University of Osnabrück through e-mail and social media. Main experimental data collected for each participant were the individual item scores and the three optional free texts. Collected data further included demographic data, English language proficiency and the order in which conditions were presented.

For each subject the responses of each item-group were averaged, which resulted in three continuous values per condition. The final C, F and S scores for each condition were computed by averaging the condition's scores from all participants. The resulting ratings of the three conditions are reported in Table 1.

Evaluation

The gathered experimental data allows answering our research questions from the introduction by checking for significant effects of the factor 'condition' (D, H and F) on the three dependent variables: S, F and C. At the same time, it appears expedient to validate the experimental setup, by checking whether the factor 'presentation order' (the 6 possible permutations in which participants might have been presented with the conditions) has, as predicted, no effect on the dependent variables. Since a within-subject design was selected, this can be done by performing one two-factor repeated measure ANOVA for each of the three variables.

Since a Mauchly Test performed on all ratings showed that the sphericity assumption is violated in the data, in the following, all ANOVA results are reported with a Greenhouse-Geisser correction.

Summary Quality

The null hypothesis regarding summary quality is that no differences exist between the quality of the three summaries.

	SS	**df**	**MS**	**F**	**P value**
Condition	96.95	1.96	49.34	100.53	2.81e−14
Order	7.41	9.82	0.75	1.54	0.16
Error	28.93	58.95	0.49		

Table 2: Greenhouse-Geisser corrected results of a two-factor ANOVA on the ratings for summary quality.

The ANOVA results in Table 2 show that 'condition' has a strongly significant effect on summary quality. A post-hoc, pairwise Tukey HSD test showed that the data-driven summary was rated significantly lower than the human or framing based summaries, whereas the latter two show no significant differences between each other (see Table 1 for the respective means). This means that the null hypothesis can be rejected, and, especially, that the framing based summary performs at human level. It is essential to put this result in the right context. The program did not summarize a natural language text at human level, where it first would have to extract and analyse the semantic content. Instead, it created a summary for a story it generated itself and for which it accordingly already possessed a computational representation of the ground truth. Also, the generated language of the summary is based on fairly rigid templates, so that any number of iterations would quickly dispel all humanoid pretensions. Notwithstanding these proper reservations, the observed performance is no trivial feat. The employed questionnaire included items regarding abstract understanding and teleology (meaning/message), which go beyond mere selective recounting.

The results also show that presentation order had no statistically significant effect on summary rating.

Framing Quality

To establish whether a better summary provides a better framing for a story, the null hypothesis can be formulated that the two conditions H and F show no significant difference in framing quality as compared to condition D. This is grounded on the previously established observation that H and F are the better summaries. Since H and F themself display no significant difference in summary quality, no prediction is made about their relationship towards each other.

	SS	df	MS	F	P value
Condition	42.43	1.94	21.88	46.37	6.11e−9
Order	5.87	9.70	0.61	1.28	0.27
Error	27.45	58.18	0.47		

Table 3: Greenhouse-Geisser corrected results of a two-factor ANOVA on the ratings for framing quality.

The ANOVA results in Table 3 show that 'condition' has a strongly significant effect on framing quality. A post-hoc, pairwise Tukey HSD test showed that all three conditions differ significantly among each other, which allows the rejection of the null hypothesis (see Table 1 for the respective means). It is interesting to note that the two summaries that were rated equally (H and F) still seem to present a differing 'frameability'. This implies that summary quality is not the only factor contributing to framing quality. The performed statistical analysis can not provide an answer to the question what these other factors might be. Here, the qualitative data collected using a free text field for each condition, asking if and how the presented text affected the participants' perception of the story, can give further insights. Its analysis can be found at the end of this section, under the heading 'Qualitative Evaluation'. For now it should suffice to say that we hypothesize that one such reason might be a phenomenological gap between summary F and readers' mental models. Summary H mainly provides coverage of the setting and events happening in the story world (contentual level), whereas summary F also analyses the actions as standing in a functional context, e.g. withholding the fruits of the protagonist's labour is described as a retaliation (artefactual level, perhaps even thematic if retaliation is taken to be the theme of the whole story). Following the assumption that thinking about a story in contentual rather than functional terms is more natural for laypeople untrained in the arts of narratological analysis, this would manifest in a phenomenological gap when reading condition F but not H. This should provide readers with a stronger impetus to re-contextualize the text, thus framing it 'as' something.

The ANOVA results again show that presentation order had no statistically significant effect on participants' ratings.

Perceived Creativity

To establish whether a better framing leads to a higher perceived creativity of the generating algorithm the null hypothesis can be formulated that all three conditions show no significant difference in creativity ratings, since the three conditions all differ in framing quality.

	SS	df	MS	F	P value
Condition	0.60	1.71	0.35	2.93	0.09
Order	1.73	8.53	0.20	1.70	0.12
Error	6.10	51.20	0.12		

Table 4: Greenhouse-Geisser corrected results of a two-factor ANOVA on the ratings for perceived creativity.

First, it should be noted that the ANOVA results in Table 4 again show that presentation order had no statistically significant effect on participants' ratings, which conclusively corroborates the choice of a within-subject design by demonstrating that subjects' judgements were not biased by previously read conditions.

The results also show that 'condition' has no significant effect on perceived creativity, which means that the null hypothesis has to be accepted. This is unexpected since, as outlined in the introduction, the field operates under the assumption that perceived creativity should benefit from framing. One explanation that would allow to uphold this assumption might be what we would call the *weak framing assumption*, which would hold that systems do benefit from framing, however only in comparison to systems that perform no framing, while differences in farming quality do not propagate on creativity ratings (which would form part only of the *strong framing assumption*). This assumption remains unfazed by the present results, since no creativity ratings were solicited without framing. However, no reason comes to mind why framing quality should be irrelevant. It should also be observed that the quality of both, summary and framing, for condition D are consistently rated very low, and that it contains a logical non sequitur regarding the mechanics of bread-baking, which cast its adequacy as framing in a doubtful light—if accepted, such a perspective would hold that condition D was essentially unframed, which then would put even the weak framing assumption under pressure.

Another avenue at interpreting this outcome is by closer scrutinizing the numerical results. It is conspicuous that all three C values are located so close to the middle of the rating scale. Such behavior was recently also observed by Riegl and Veale (2018), who interpreted it as a symptom of participants' boredom or overtaxation. Beyond questioning the data, only a closer look at the item-based breakdown of the question group 'creativity' remains (see Table 5). While this might aid in satisfying one's curiosity, it should be clear that any analysis searching for significant difference on an item-based level remains prohibitive, because it would constitute a retesting of the same data, and be thus prone to false positives.

Considering the individual items C1 through C6 in Table 1 it becomes clear that a summary-based framing can not be expected to contribute equally to the individual ratings. The typicality (C2) and the novelty (C3) of a story are less likely to be significantly increased just by the merit of a short summary. On the other hand, a fitting but unexpected summary can well be taken as an indication for a

	interesting	typical	novel	imaginative	appreciative	skillfull
D	2.94	3.58	2.28	3.06	2.69	3.47
F	3.17	3.72	2.44	3.28	3.06	3.39
H	3.17	3.89	2.28	3.08	3.08	3.36

Table 5: Item-based breakdown of the question group 'creativity', reporting the condition-wide means for the items C1 through C6 (labeled with the quality they intend to capture for better readability).

system's ability for appreciation of its own product (C5). Indeed, the appreciation scores seem to suggest an effect of the two 'good-summary conditions'. As mentioned, a statistical test of this effect can unfortunately not be conducted. Thus, questions remain.

Qualitative Evaluation

The free text data collected from participants in order to understand how the different summaries might have affected their perception of the story can be used to analyse why conditions F and H differ in framing quality, while there are no significant differences in the respective summary qualities. A first corroboration of this statistical result in the introspective data is the fact that this voluntary field contained 12 relevant[7] answers for condition F, while only 6 for H. This implies that subjects' perception of the story was stronger affected by condition F, which is an indicator of more effective framing.

To analyse how subjects' comments differ between the two conditions, two coders were employed to code all of the 18 comments. Their task was: "For each description, select one to three categories, which best describe what aspect of the reader's perception was changed by, or at least was different in, the summary". The categories available as codes were explained as follows:

- *function*: The text describes a change in perception of the function of certain events for the story as a whole. This includes the judgement that events form part of a high-level structural unit, like a 'hero's journey', are fulfilling a narrative function like 'introducing a conflict', or take on an unexpected meaning like 'deceiving an opponent'.
- *character*: The text describes a change in perception of characters or their interrelation. This can include individual's motivations, emotions or reasoning, their perceived personality as well as attitudes towards each other.
- *theme*: The text describes a change in perception of the story's moral (example moral: 'don't stray from the right path'), or which abstract themes of the human condition it represents (example: 'search for the meaning of life').
- *other*: everything that doesn't fit the above categories.

Since any of the texts can contain commentary on several of these aspects, the results were interpreted as a one-to-many classification, for which recently a Cohen's kappa-like measure of inter-coder agreement called Fuzzy Kappa was introduced (Kirilenko and Stepchenkova 2016). In our data, fuzzy unweighted kappa between the two coders is 0.60, which according to Landis and Koch (1977) is the border between moderate and substantial agreement. The aggregated results of the two codings of subjects' comments can be found in Table 6, which depicts which percentage of codes were of which type in the two conditions. The most marked difference between condition F and H can be observed in the proportion of codes of the type 'functional'. Both coders determined that the latter condition did not affect subjects' perception of what certain events meant for the plot, while in the former condition around 30% of the indicators of differing meaning referred to this category. The other three codes do not yield such conclusive differences.

Our hypothesis, already outlined above, is that the functional perspective taken in condition F is uncommon to lay readers and for this reason works as a framing. However the analysis here is only a further indication for our case, and should be best read as a correlation: the summary that affected subjects' perception of the story on a functional level happens to be the summary that is judged to be the better framing. To prove causation, more study would be required.

		character	function	theme	other
coder 1	F	0.56	0.31	0.13	0.00
	H	0.50	0.00	0.25	0.25
coder 2	F	0.53	0.27	0.13	0.07
	H	0.50	0.00	0.17	0.33

Table 6: Distribution of codes per coder and condition, as relative frequencies. Each row represents an overview of how often subjects' perception of the story was affected in a specific domain (i.e. code) by the respective condition's text.

Conclusion

The study presented in this paper has shown three things. First, it has demonstrated that an FU-Analysis based approach to story summarization can perform at human level—if employed on top of a plot generation system that implements the phenomena required to model functional units. By comparing the suitability of summaries of different quality for framing stories it, secondly, has shown that a better summary is also a better framing. This opens up an interesting avenue for storytelling systems to perform summary-based framing of generated artefacts, a further step up climbing the meta-mountain (Colton and Wiggins 2012). However, this is not the whole story, as summaries of comparable quality have shown differing suitability for framing. Our tenta-

[7]This count excludes answers like "It had no effect on my perception", which were filtered out before all further analysis.

tive advise to researchers interested in employing summary-based framing is to aim for creating a phenomenological gap between the level of abstraction at which consumers and the system reason about the plot. One possible approach to achieve this is FU-Analysis based summarization. More research is needed into how to generate summaries at other levels of abstraction, like for instance the thematic level concerned with higher meaning, symbolism or messages. Interestingly, the morals that Minstrel (Turner 1993) provided for its knight stories can be seen as one instance of framing based on thematic summary, developed long before the term framing itself was coined. Third, the study failed to show an effect of framing on the perceived creativity of a computational system. So far, it remains unclear to us whether this is due to the complex design of our study, an overly general creativity questionnaire or, perhaps, to the fact that framing actually doesn't work. Thus, we invite researchers with systems that practice different kinds of framing, or framing in a different domain than narrative, to explore whether they can replicate these results.

Acknowledgment.

We are grateful to Dr. Annette Hohenberger from University Osnabrück for her guidance in creating the questionnaire, and to Sven Wilke for the underlying implementation. This work was funded by an Alexander von Humboldt Ph.D. fellowship.

References

Berov, L., and Kühnberger, K.-U. 2018. An evaluation of perceived personality in fictional characters generated by affective simulation. In *Proceedings of the Ninth International Conference on Computational Creativity*, 24–31. Salamanca, Spain: Association for Computational Creativity.

Berov, L. 2017. Steering plot through personality and affect: an extended bdi model of fictional characters. In *KI 2017: Advances in Artificial Intelligence: Proceedings of the 40th Annual German Conference on AI*, 293–299. Cham: Springer.

Charnley, J. W.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *Proceedings of the 3rd International Conference on Computational Creativity*, 77–81.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: the final frontier? In *Proceedings of the 20th ECAI*, volume 12, 21–26. IOS Press.

Colton, S.; Charnley, J. W.; and Pease, A. 2011. Computational creativity theory: The face and idea descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, 14–20.

Colton, S. 2012. Evolving a library of artistic scene descriptors. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design*. Springer. 35–47.

Eder, J. 2010. Understanding characters. *Projections* 4(1):16–40.

Gambhir, M., and Gupta, V. 2017. Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.* 47(1):1–66.

Goyal, A.; Riloff, E.; et al. 2013. A computational model for plot units. *Comput. Intell.* 29(3):466–488.

Kirilenko, A. P., and Stepchenkova, S. 2016. Inter-coder agreement in one-to-many classification: fuzzy kappa. *PloS One* 11(3):e0149787.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 159–174.

Lehnert, W. G. 1981. Plot units and narrative summarization. *Cogn. Sci.* 5(4):293–331.

Norton, L. 1917. The richard mutt case. *The Blind Man* (Vol. 2):5.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Riegl, S., and Veale, T. 2018. Live, die, evaluate, repeat: Do-over simulation in the generation of coherent episodic stories. In *Proceedings of the Ninth International Conference on Computational Creativity*, 80–87. Salamanca, Spain: Association for Computational Creativity.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds Mach.* 17(1):67–99.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Turner, S. R. 1993. *Minstrel: a computer model of creativity and storytelling*. Ph.D. Dissertation, University of California at Los Angeles, Los Angeles, CA.

Wilke, S., and Berov, L. 2018. Functional unit analysis: Framing and aesthetics for computational storytelling. In *Proceedings of the 7th International Workshop on Computational Creativity, Concept Invention, and General Intelligence*, volume 2347 of *CEUR Workshop Proceedings*.