How Narratives Move Your Mind: A Corpus of Shared-Character Stories for Connecting Emotional Flow and Interestingness

Yusuke Mori[a,*], Hiroaki Yamane[b], Yoshitaka Ushiku[a], Tatsuya Harada[a,b]

[a] *The University of Tokyo*
[b] *RIKEN*

Abstract

Creativity is considered a human characteristic; creative endeavors, including automatic story generation, have been a major challenge for artificial intelligences. To understand how humans create and evaluate stories, we (1) construct a story dataset and (2) analyze the relationship between emotions and story interestingness. Given that understanding how to move readers emotionally is a crucial creative technique, we focus on the role of emotions in evaluating reader satisfaction. Although conventional research has highlighted emotions read from a text, we hypothesize that readers' emotions do not necessarily coincide with those of the characters. The story dataset created for this study describes situations surrounding two characters. Crowdsourced volunteers label stories with the emotions of the two characters and those of readers; we then empirically analyze the relationship between emotions and interestingness. The results show that a story's score has a stronger relationship to the readers' emotions than the characters' emotions.

Keywords: Natural Language Understanding, Story Evaluation, Emotion, Creativity, Text Corpus Construction, Crowdsourcing

[*] This is to indicate the corresponding author.
Email address: mori@mi.t.u-tokyo.ac.jp (Yusuke Mori)
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan

1. Introduction

Since ancient times, humans have enjoyed both reading and creating stories. Many studies have aimed to understand how captivating stories are created—for example, by analyzing common structures in myths and folktales [1, 2]. Moreover, these studies have been used as reference for those creating new works. In recent years, practical story creation techniques [3, 4] have been devised and employed to compose stories that attract attention.

Because creativity is considered to be a human characteristic, many researchers have been interested in how it might be automated using computers. In recent years, research has been conducted on story generation [5, 6, 7] and reading comprehension of stories [8, 9, 10, 11]. To automate these creative activities, computers should know what kinds of stories draw readers in.

In this study, to understand how humans create and evaluate amazing stories, we examined how short stories satisfy readers. In story and narrative research, it is necessary to first define a story and the kind of text that can be regarded as a story; the task of judging whether a text is a story is known as story detection [10]. We define a story as a series of events related to characters and having a beginning and an end; these events are intended to change the emotions and relationships of the characters.

However, a major challenge remains in story generation: there is no established standard for a good story. Indices for evaluating reader satisfaction must be established to determent whether a story that has been generated satisfies readers [12]; however, it is difficult to fairly compare stories of different lengths and genres. Moreover, a variety of elements affect the "interestingness" of a story—even when different stories are created from the same idea, the interestingness of the completed stories differs depending on how the writers expanded the initial idea.

Therefore, we first consider evaluating reader interest in short stories of equal length that can be created in one sitting. Our focus is story expansion, rather than developing a system that creates entire stories. In this study, we assume

Title: Friends Never Die

Story: Thomas was about to play piano at a local concert hall. He was to play a piece written by his close friend Linda. She had died earlier that year of cancer. As he started to play he hesitated and couldn't bring his hands to play. He looked to the audience and couldn't believe his eyes: it was Linda smiling at him and urging him to go on and he did beautifully bringing the audience to their feet.

Settings: (1) Thomas is a pianist. (2) Linda is a close friend of Thomas.

Evaluation Score (ranging from 1 to 5, where 1 is bad and 5 is good):

Total Score: 4.57, Storyness: 4.14, Frequency: 4.00, Consistency: 4.00, Clarity: 4.43, Meaning: 4.57

Reviews:

- The story was interesting and meaningful. It was kind of sad but had a happy ending.

- few memories of special persons dont simply fade away

- Even after death she helped him

- The story explains about the friendship even after death.

- It was interesting for a number of reasons. That he was playing a piece written by a friend who died; that he had trouble playing. The downside was the ending was ridiculous which is why I didn't rate it higher.

- I feel so sad but proud of you guys.

- This story explains about friendship between Thomas and Linda who are pianists. It is really amazing explain a close friend's feelings when his close friend is died.

Table 1: An example of a human-written story that was evaluated as good. The evaluation item scores and total score are the mean ratings of seven volunteers.

that it is useful to consider the "goodness" of a short story when measuring the goodness of a long story that is expanded from it. We therefore focus on how evaluation scores improve when we expand a story. For fair comparison, we propose a "shared-character story" method, wherein character settings are shared before the stories are written. We thus created a novel story dataset. Table 1 shows an example collected story.

For part of this dataset, crowdsourced volunteers annotated the stories with their degree of satisfaction as readers. In addition, the emotions conveyed and evoked in each sentence were also labeled by the readers. Practical creative techniques for satisfying readers stress the importance of being conscious of readers' emotions. There have also been attempts to classify stories by drawing their emotional arcs [13]. Therefore, we investigated whether the emotional flow of a story is useful for predicting reader interest. Earlier research has involved annotating only one aspect of emotion; in contrast, we noted the emotions of each character and of the readers. The stories we collected always had two main characters so that emotions could be judged similarly across stories. Further, to evaluate various emotions, we referred to Russell's circular model (shown in Figure 1) and used emotions expressed on two axes: ***Valence*** (positive/negative) and ***Arousal*** (excited/calm) [14]. Figure 2 shows the emotional flow of the story in Table 1.

This study makes two major contributions:

Creation of a Story Dataset: We propose a method of generating a story dataset in which character settings are shared before the short stories are created. We created this dataset through crowdsourcing.

Analyses of Emotional Flow and Interestingness: Using a subset of this dataset, we had crowdsourcing volunteers annotate a story interestingness, write review comments, evaluate the stories from multiple perspectives, and note the emotions in each sentence. Using the annotated dataset, we analyzed the relationship between stories' emotions and interestingness and found that the interestingness of a story is more related to its readers' emotions than the character emotions predicted from the story.

2. Related Work

2.1. Story Dataset Studies

Several methods have been proposed for determining whether machine learning models can read and understand stories [15, 16, 17, 18, 19]. The *Story Cloze Test* requires models to judge which of two choices is a correct ending for a set

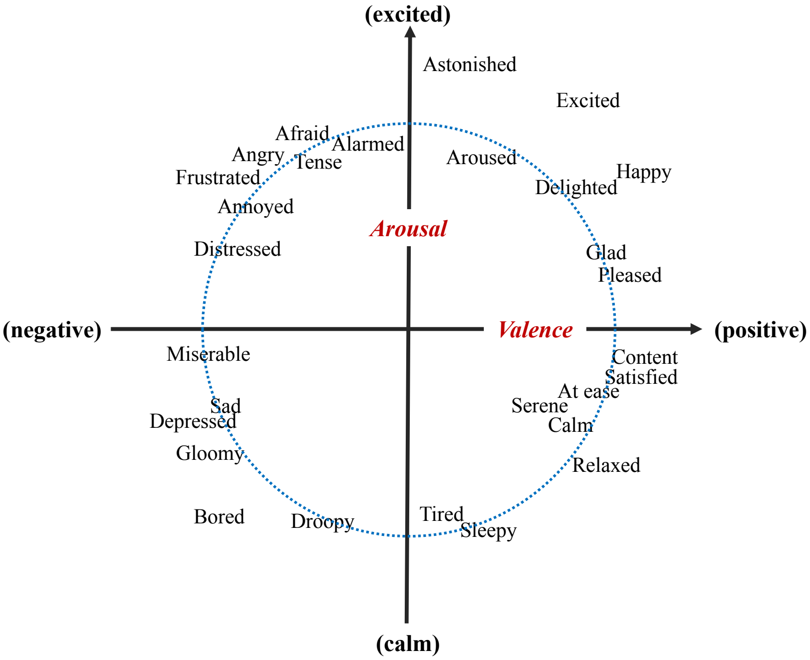


Figure 1: The circular model of emotions expressed on two axes: *valence* and *arousal*. The positions of the 28 emotions are approximated from the original paper.

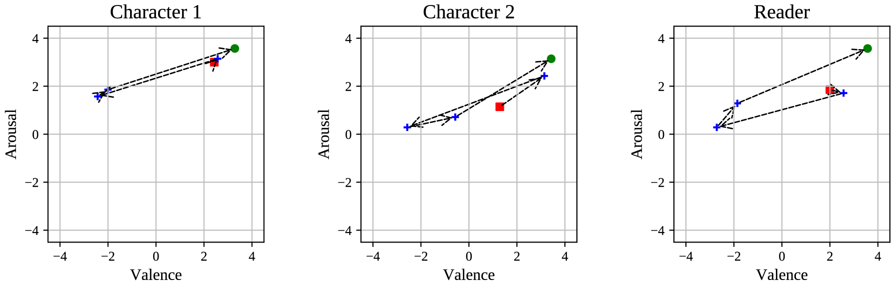


Figure 2: The emotional flow of the story shown in Table 1. The red square indicates the starting point (the first sentence) and the green circle indicates the end (the last sentence). The emotional movements are indicated by black arrows.

of consistent sentences [11, 20]. A story dataset called *ROCStories*, consisting of five-sentence-long short stories collected by crowdsourcing on Amazon's Mechanical Turk (MTurk), was proposed for use with this task. These researchers chose to collect five-sentence-long stories because they claimed that this length is sufficient to establish a story; any increase in length would introduce more trivial matters. In contrast, *ROCStories* did not give writers any theme prompts and had them write freely. To clarify the association between conditions and products, we shared character settings with the writers and asked that they write stories based on these settings.

2.2. Studies of Emotion in Story

Research on emotion and text strives to better understand human-written texts [21, 22]. In the domain of story or narrative, Chaturvedi et al. [9] proved that by considering emotional movement in a story, models can improve their performance on the *Story Cloze Test*. Studies have previously investigated the relationship between emotion and story interestingness. Reagan et al. [13] showed that stories collected from Project Gutenberg could be classified into six styles by considering their emotional arcs (i.e., the trajectory of average happiness in a story).[2] Alm et al. [8] provided children's novels with positive/negative sentiment evaluations and with a more complex set of eight classes of emotions based on Ekman's basic emotions [23]. To predict emotions using machine learning, they classified the eight emotions again into three emotional valences: positive, negative, and neutral. In this study, to consider not only positivity/negativity, but also more complicated emotions, we use two annotation axes—valence and arousal—based on Russell's circular model [14].

2.3. Story Evaluation

To evaluate the goodness of a story, researchers have proposed the approach of collecting story-like texts from social networks [12]. Story detection was

[2] http://www.gutenberg.org/

performed via crowdsourcing to create a large dataset. The number of upvotes the text received on the social networking platform was used to indicate the goodness of the story. Story quality was therefore defined as the number of people interested in the story. Here, we assume that a story's quality is related to reader satisfaction. The difference in our approach from the previous study is that we collected texts that were intended by their writers to be stories, and gathered and analyzed scores from readers who approached the task from the perspective of story evaluation.

3. Dataset Construction

In Section 3.1, we present how we created two-sentence settings to be shared. In Section 3.2, we describe our novel approach to story collection based on the prepared settings. In Section 3.3, we explain how we assign scores and emotions to the collected stories. We created two-sentence settings about two characters and their relationship. We sought shared-character stories, wherein each writer wrote independently but following the same shared character settings. With the assistance of MTurk volunteers, we obtained a dataset consisting of 759 short stories based on the character settings we provided. For part of the constructed story dataset, we conducted another MTurk task in which the stories were annotated with scores and emotions.

3.1. Shared Character Settings

We referred to the method with which *ROCStories* collected short stories [11] and also collected our stories on MTurk.[3] However, we took the additional step of preparing basic character settings that the volunteer short story writers were required to follow. The task was designed to permit multiple volunteers to work on the same setting.

The two-sentence settings given to the volunteers were formatted to give (A) the name and job of the first character and (B) the name of the second character

[3] https://www.mturk.com/

and the relationship between the two characters. As defined above, a story is a series of events that change the emotions of and relationships between the characters. For simplicity, we provided settings containing only two characters. To limit situations and create a common starting point, we set a job for the first character.

For creating two-character settings, we first create the name and job of the first character. The setting of a job was to recall everyday characters such as company employees, students, and athletes. Then, we create the second character with his/her relationship to the first character. The second setting includes various cases such as when the second character is a family member or a colleague of the first character, when the relationship of the two is positive or negative, or even when the second character is not a human. In Table 2, we show some examples of the shared character settings. Every time a volunteer participated in our story writing task, the volunteer was given a two-sentence setting (for example, (1) "Barbara is a teacher." and (2) "Robert is Barbara's brother.") and wrote a story based on the setting.

Similar to the "shared world" method wherein each work is created by sharing the same story setting, we sought shared-character stories, wherein each writer wrote independently but following the same shared character settings.

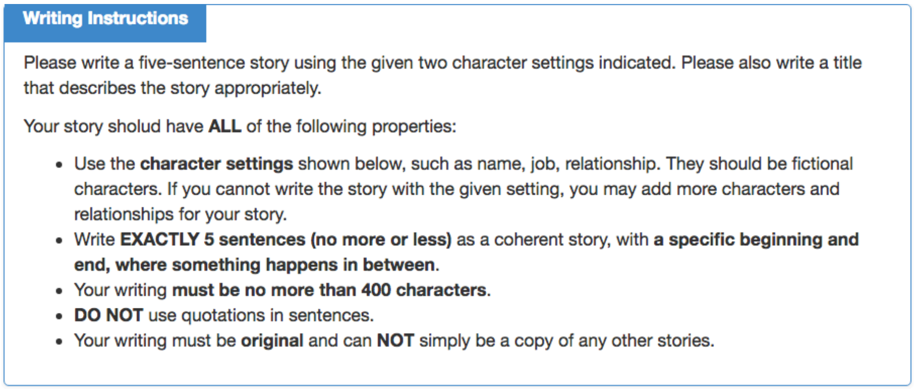
3.2. Story Collection

Using the two-sentence settings prepared in Section 3.1, we constructed our novel dataset of shared-character stories with the assistance of MTurk volunteers.

When evaluating a story's interestingness, the amount of information available varies with the number of sentences in the story. To evaluate each story under equivalent conditions, we assumed that it would be necessary for stories to comprise the same number of sentences. With reference to *ROCStories*, when collecting stories to create a new dataset, we required volunteers to write exactly five sentences, which is relatively short but is sufficient to construct a storyline [11]. If the sentence length exceeds this number, the effect of writing style on

(A) the name and job of the first character	(B) the name of the second character and the relationship between the two characters
- Barbara is a teacher.	- Robert is Barbara’s brother.
	- Robert is Barbara’s colleague.
	- Robert is Barbara’s rival.
	- David is a close friend of Barbara.
	- David is a student of Barbara.
	- David is a person that Barbara does not like.
	- Dorothy is Barbara’s sister.
	- Dorothy is Barbara’s colleague.
	- Dorothy is Barbara’s rival.
	- Linda is a close friend of Barbara.
	- Linda is a student of Barbara.
	- Linda is a person that Barbara does not like.
	- Max is Barbara’s dog.
	- Max is a dog owned by a friend of Barbara.
	- Max is a respected person of Barbara.
- Kevin is an office worker.	- Robert is Kevin’s brother.
	- Robert is Kevin’s colleague.
	- Robert is Kevin’s rival.
	- David is a close friend of Kevin.
	- David is Kevin’s boss.
	- David is a person who Kevin does not like.
	- Dorothy is Kevin’s sister.
	- Dorothy is Kevin’s colleague.
	- Dorothy is Kevin’s rival.
	- Linda is a close friend of Kevin.
	- Linda is Kevin’s boss.
	- Linda is a person who Kevin does not like.
	- Max is Kevin’s dog.
	- Max is a dog owned by a friend of Kevin.
	- Max is a respected person of Kevin.

Table 2: Examples of provided two-character settings. We first created the name and job of the first character (A), then create the name of the second character and the relationship between the two characters (B).



Writing Instructions

Please write a five-sentence story using the given two character settings indicated. Please also write a title that describes the story appropriately.

Your story sholud have **ALL** of the following properties:

- Use the **character settings** shown below, such as name, job, relationship. They should be fictional characters. If you cannot write the story with the given setting, you may add more characters and relationships for your story.
- Write **EXACTLY 5 sentences (no more or less)** as a coherent story, with **a specific beginning and end, where something happens in between**.
- Your writing **must be no more than 400 characters**.
- **DO NOT** use quotations in sentences.
- Your writing must be **original** and can **NOT** simply be a copy of any other stories.

Figure 3: The instruction snippet used in our MTurk story writing task.

interestingness can become overemphasized and the effect of the storyline on interestingness is underestimated. To discover the relationship between storylines and emotional flow, we therefore set the story length to five sentences.

The following additional conditions were included in the instructions: (a) A story should have a clear start point and end point. (b) A story should not include quotation marks (to avoid dialog). (c) The number of sentences is limited to prevent the story from straying from its main point. (d) Writing must be original and cannot simply copy another story.

In Figure 3, we show the instruction snippet for our story writing task. The tasks were issued multiple times, so that adjustments could be made to ensure that our instructions were clearly understood. The figure shows the final set of instructions. For example, we initially received answers in which writers composed longer stories, not observing the condition to “Write 5 sentences”; hence, we added the word “EXACTLY” to this instruction to better convey the desired condition. We also restricted the number of characters that should be used, but we considered this to be an auxiliary measure and did not strictly apply it. Since we wanted writers to write a story as freely as they could, we did not state that the stories would be evaluated with a focus on emotions.

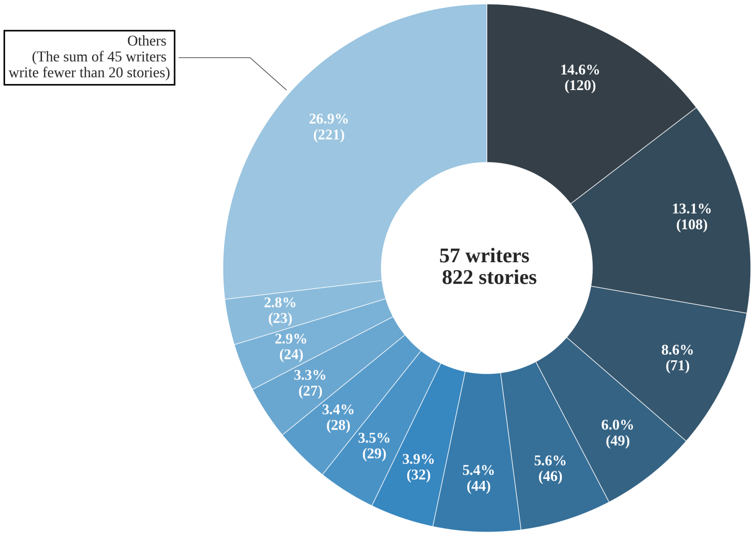


Figure 4: The distribution of the number of stories per writer as a pie chart. For the 12 writers who wrote more than 20 stories each, the number of stories written and the percentage of the dataset corresponding to those numbers are shown. Writers who wrote fewer than 20 stories were compiled into the category of "others."

We had four volunteers write about the same character settings.[4] We collected 822 stories in total and excluded stories that did not contain five sentences; thus, our final dataset contained 759 stories.

In Figure 4, we show the distribution of the number of the number of stories written per MTurk worker as a pie chart. Our 822 collected stories were written by 57 writers; the writer who wrote the most authored 120 stories (14.6% of the dataset). The dataset has diverse stories written by many writers. We should note that this analysis is done for all the 822 collections, not for the 759 stories we finally obtained.

In the 759 stories that we obtained, there were 214 combinations of two-

[4] There are settings used for task design and test. For these settings, we had less than four volunteers.

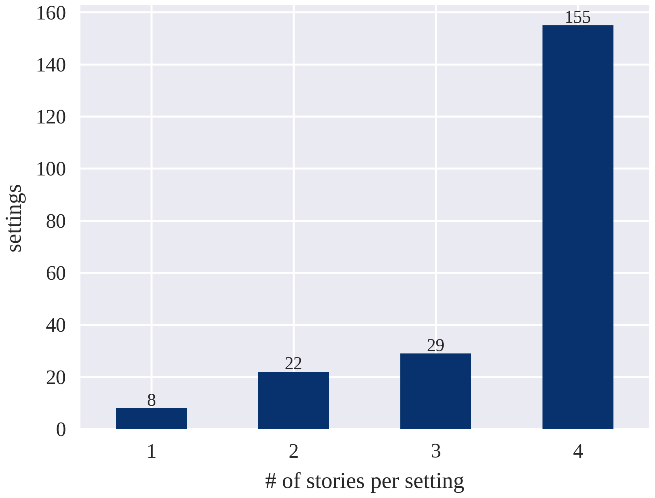


Figure 5: The number of stories per setting.

character settings. The number of stories per setting are shown in Figure 5, which shows that most of the settings were used to write 4 stories. Note that the combination of the statements (1) "Mary is a high school student." and (2) "Linda is a close friend of Mary." combine to form one "setting" in this bar graph.

Examples of the collected stories and their corresponding character settings are shown in Table 3. Table 4 shows four different stories written in the same character setting.

We call the dataset created in this study our set of shared-character stories.[5]

Our shared-character stories consider how volunteers expand stories from the same settings. The given settings should enable volunteers to generate stories easily. For that reason, we use settings reminiscent of everyday life. As writing stories is a nearly unconstrained activity, giving all possible settings and collecting all possible stories is not realistic. To check that the stories generated

[5] The dataset of shared-character stories created in this study will be publicly available in the future.

Settings	**Title**	**Story**
(1) Margaret is an athlete. (2) David is a friend of Margaret.	The New Competitive Diver	Margaret had been diving for years. Her friend David encouraged her to dive in competitions. Margaret was scared. David had dived in competitions before. He reassured her that she would do great.
(1) Kevin is an office worker. (2) Dorothy is Kevin's sister.	Kind brother	Kevin is a officer at a big corporate firm. Dorothy is Kevin's sister. Dorothy once needed $5000 for her college fees. Kevin helped Dorothy with financially. Dorothy is always grateful to Kevin for his help.

Table 3: Examples of collected stories and their corresponding character settings. In the two settings, we gave volunteers a character with a job that recalls everyday life and another character with a relationship to the first. Note that "athlete" was misspelled as "athelete" in our settings, but no workers seemed to be confused by the typo.

from our settings are not biased about storylines, we compared the distribution of the shared-character stories with that of *ROCStories*, which consists of stories written by volunteers without any settings given. Here, we define a story's distribution as the distributions of words therein.

The frequency of word appearance in our shared-character stories is shown in Figure 6, which indicates the 30 most frequent words appearing in the body of the 759 stories. Stop words and characters' names were excluded. We used English stop words list contained in the NLTK toolkit [24], and added ".", ",", "'s", "!", "?", "n't" to our stop words list. All words were changed to lowercase and to lemmas. Trends in storylines can be examined from this word frequency. Since words such as "school," "day," and "work" frequently appeared, we can say that most of the stories were about everyday life, as we intended. Figure 7 shows that about half of the 30 most frequent words appearing in *ROCStories* are the same as the most common words in our story dataset, suggesting that we collected nonbiased stories written by volunteers.

Title	**Story**
High School Nerves	Mary was nervous her first day of high school. It was going to be so much different than middle school; everyone was so much more grown up. Some of the boys even had beards! The only thing that made her feel better was that her very good friend Linda had a few of the same classes as she did. At least she was guaranteed one friendly face!
friendship prevails	Mary and Linda were together in high school. They had been friends since kindergarten. Their friendship only grew with time. When Mary broke up with her boyfriend, it was Linda who stayed by her side. Mary considered herself lucky to have such a good friend.
Friends should never fight.	Mary and Linda have been good friends since elementary school. They have sleep overs and have study groups. However, since Mary got her licence she has been ignoring Linda. Linda is always asking her for rides places and Mary feels Linda is just using her. Mary told Linda who she feels and the girls made up and are now friends again.
My Lost Love	Mary and Linda are close friends who attend the same High School. One day while studying, they kiss and realize they have feelings for one another. Linda is a very Conservative Christian, so she hides the relationship. Linda's mother catches Mary and Linda kissing and sends Linda away to boarding school. Linda returns home years later to find out that the love her her life, Mary, has passed away.

Table 4: Example of stories written with the same setting. The correspondent setting was (1) "Mary is a high school student." (2) "Linda is a close friend of Mary." All four stories written in the setting were about a high school student and her close friend, but how the story was expanded differed with each author.

ROCStories emphasized a story's consistency rather than its entertainment value or the drama it contained, and gathered nonfictional stories; however, fictional stories are suitable for our task. *ROCStories* were collected on the assumption that they are used for reading comprehension tasks and handling

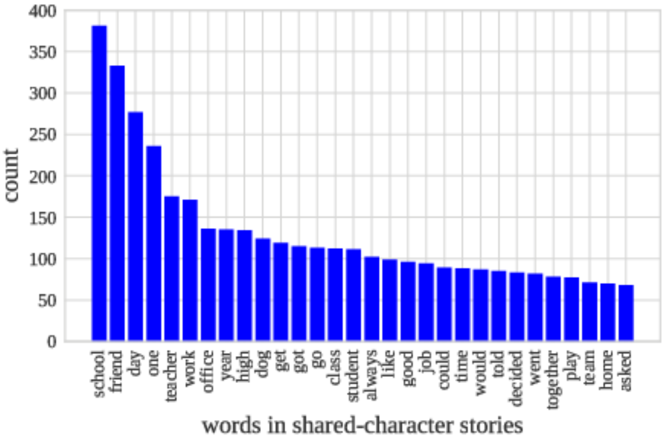


Figure 6: The frequency of words appearing in the bodies of the collected stories. Stop words and characters' names were excluded. All words were changed to lowercase and to lemmas. The 30 most frequent words are shown.

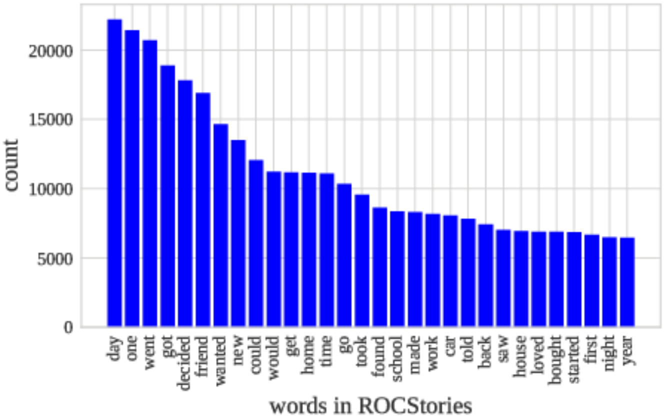


Figure 7: The frequency of words appearing in the bodies of the *ROCStories*. Stop words were excluded. All words were converted to lowercase and replaced with their lemmas. The 30 most frequent words are shown. This graph has 16 words in common with the graph in Figure 6: "went," "day," "get," "work," "told," "would," "year," "decided," "could," "got," "school," "time," "go," "friend," "one," and "home". Additionally, comparing the 100 most frequent words in each dataset, we found that 60 words occurred in both lists.

common sense. In contrast, we are concerned with story generation, which places greater importance on the included stories being evaluated as "interesting."

3.3. Annotation of Scores and Emotions

Using a subset of our shared-character stories, we created a dataset consisting of 100 stories and annotated them via crowdsourcing. The dataset for annotation consisted of 75 human-written stories and 25 artificially modified stories.[6] The breakdown of the annotation dataset is as follows:

Human written: Human-written stories randomly taken from the shared-character story dataset (75 stories).

Random last sentence: A randomly chosen story from the shared-character stories dataset with its fifth sentence replaced by the fifth sentence of another story, also randomly selected from shared-character stories (10 stories). To create these stories, we randomly extracted the required number of stories from the shared-character stories dataset and then separately extracted the same number of endings from the dataset and replaced them with the endings of the original stories.

Random last sentence from the same setting: Like the previous group, but the last sentence was required to come from another story with the same setting as the first (10 stories). Stories that did not share their setting with any other story (there were eight such stories, see Figure 5) were not used for this case.

Random order: A story randomly selected from the shared-character stories, with its five sentences rearranged (5 stories).

We compared the artificially modified stories with human-written stories to confirm that the stories included in our dataset were sufficiently good. We focused on modifying the last sentence in these stories because humans can

[6] To ensure a fair evaluation, we decided not to exclude stories which may contain offensive content or expressions. Whether a representation is offensive depends on cultural and personal values; in this task, unpleasantness should be judged by the volunteer as a reader, and not by us. We obtained consent from the annotation volunteers to show potentially offensive content and expressions in the stories.

choose the correct ending from two choices (correct and wrong) in a well-written story with high accuracy [11]. Using this dataset for annotation, we carried out additional crowdsourcing using MTurk. We had the volunteers assign a total score (i.e., story interestingness) to each story.

We also had the volunteers write a review comment for each story, as we believe that reviews contain a lot of information that can be extracted to obtain a deeper understanding of the story evaluations.

Volunteers were required to evaluate each story not only in terms of general interest, but also in terms of the five evaluation aspects below. Stories were scored on a five-point scale.

Storyness: Does the text seem to be a story?

Fluency: Does the story read smoothly and fluently?

Consistency: Is the story coherent from sentence to sentence?

Clarity: Is the content of the stories easy to understand?

Meaning: Does the story have a meaning/message?

Moreover, we focused on the role of emotion as an index for evaluating reader satisfaction. We required volunteers to annotate emotions for each sentence in the story, to investigate the relationship between story interestingness and the emotions of the characters and the reader. The novelty of our approach for codifying the emotions in a text is that we considered emotions occurring from different viewpoints. The emotions of each character and the emotions felt by the reader were considered separately. For example, the main character and characters opposing him/her should have conflicting emotions. When the hero faces a challenge, the hero himself may be depressed, but the reader may be excited, expecting a counterattack by the hero, or angered by the tragedy striking the hero. To consider these differences, we had the volunteers think about the emotions of “Character 1” (the person indicated in the first part of the setting), “Character 2” (the person indicated in the second part of the setting), and the “Reader” (the volunteer).

Evaluation criteria with which volunteers could annotate emotions had yet to be established. We decided to evaluate emotions on two axes, valence and arousal. For this, we referred to Russell's circular model [14], which states that complex emotions can be expressed using two axes, valence (positive/negative) and arousal (excited/calm). Furthermore, based on previous studies of self-reported emotions [25, 26], the volunteers in this study were required to report their valence and arousal on a nine-point scale for each axis.

The contents of the task are summarized as follows:

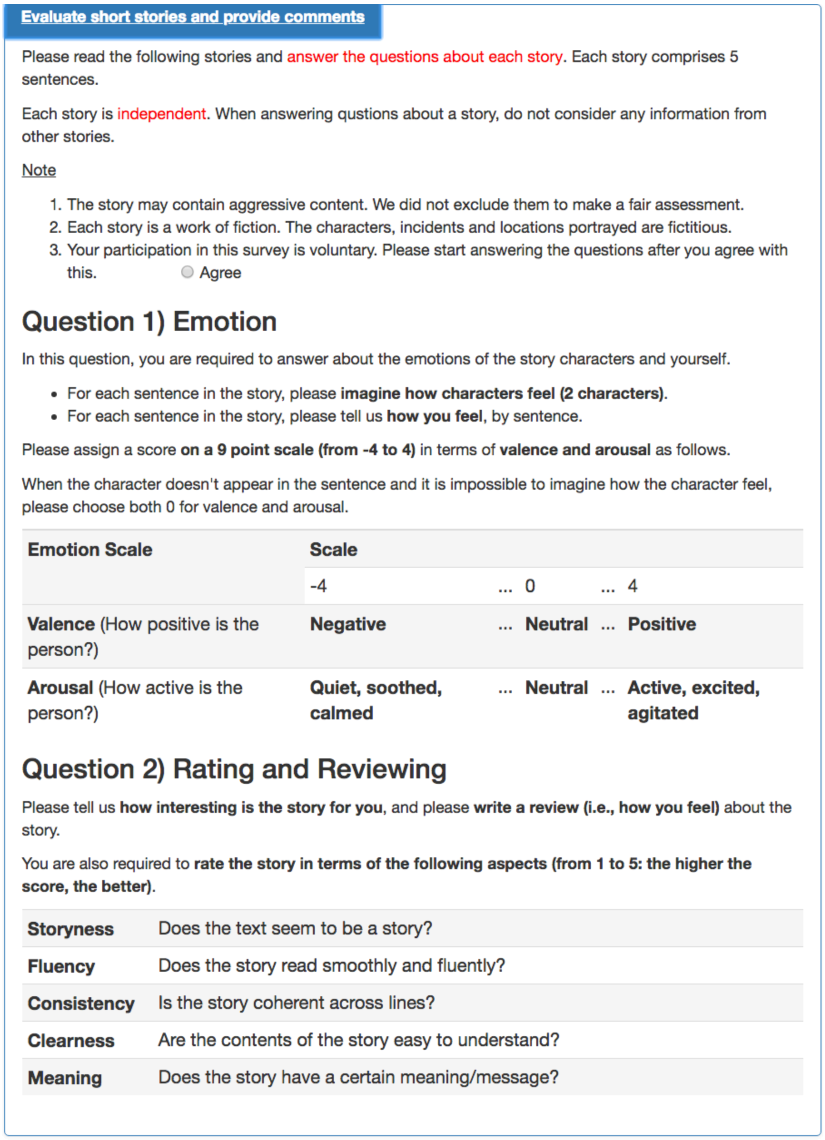
- Assign emotion values for each sentence from the points of view of Character 1, Character 2, and the Reader. Two axes of emotions were evaluated. Volunteers scored valence and arousal on a nine-point scale (ranging from -4 to 4) for each story.
- Assign a total score to the story on a five-point scale (ranging from 1 to 5) and write a review of the story.
- Assign scores on a five-point scale (ranging from 1 to 5) for each evaluation aspect: *Storyness*, *Fluency*, *Consistency*, *Clarity*, and *Meaning*.

We show in Figure 8 the instruction snippet for our MTurk story evaluation task. When a character does not appear in the sentence and it is impossible to imagine how the character feels, we had workers choose 0 for both valence and arousal.

We had 33 volunteers annotate each of the 100 stories in the dataset, obtaining 623 answers in total. Our constructed dataset may seem small, but studies have been able to offer considerable new insight into text understanding with relatively small datasets [8, 27]. We believe the dataset is large enough to begin considering this new research topic regarding stories and emotions.

In Figure 9, we show the distribution of the number of annotations. Thirty-three readers produced 623 annotations; the reader who annotated the most submitted 81 annotations (13.0%).

To investigate the demographic of MTurk volunteers who cooperated with our story evaluation task, we conducted an MTurk survey to gather the characteristics of the 33 readers. We asked readers about their gender, age, location



Evaluate short stories and provide comments

Please read the following stories and answer the questions about each story. Each story comprises 5 sentences.

Each story is independent. When answering qustions about a story, do not consider any information from other stories.

Note

1. The story may contain aggressive content. We did not exclude them to make a fair assessment.
2. Each story is a work of fiction. The characters, incidents and locations portrayed are fictitious.
3. Your participation in this survey is voluntary. Please start answering the questions after you agree with this. ○ Agree

Question 1) Emotion

In this question, you are required to answer about the emotions of the story characters and yourself.

- For each sentence in the story, please **imagine how characters feel (2 characters)**.
- For each sentence in the story, please tell us **how you feel**, by sentence.

Please assign a score **on a 9 point scale (from -4 to 4)** in terms of **valence and arousal** as follows.

When the character doesn't appear in the sentence and it is impossible to imagine how the character feel, please choose both 0 for valence and arousal.

Emotion Scale	**Scale**				
	-4	...	0	...	4
Valence (How positive is the person?)	**Negative**	...	**Neutral**	...	**Positive**
Arousal (How active is the person?)	**Quiet, soothed, calmed**	...	**Neutral**	...	**Active, excited, agitated**

Question 2) Rating and Reviewing

Please tell us **how interesting is the story for you**, and please **write a review (i.e., how you feel)** about the story.

You are also required to **rate the story in terms of the following aspects (from 1 to 5: the higher the score, the better)**.

Storyness	Does the text seem to be a story?
Fluency	Does the story read smoothly and fluently?
Consistency	Is the story coherent across lines?
Clearness	Are the contents of the story easy to understand?
Meaning	Does the story have a certain meaning/message?

Figure 8: The MTurk instruction snippet for our story evaluation task. Although we used the term "Clearness" in the instruction, for better understanding we replaced it with "Clarity" in this paper.

(country), language skills, educational background, and reading preferences. We did not ask for identifying personal information (e.g., address) and the information collected was anonymized and used so that participants could not be

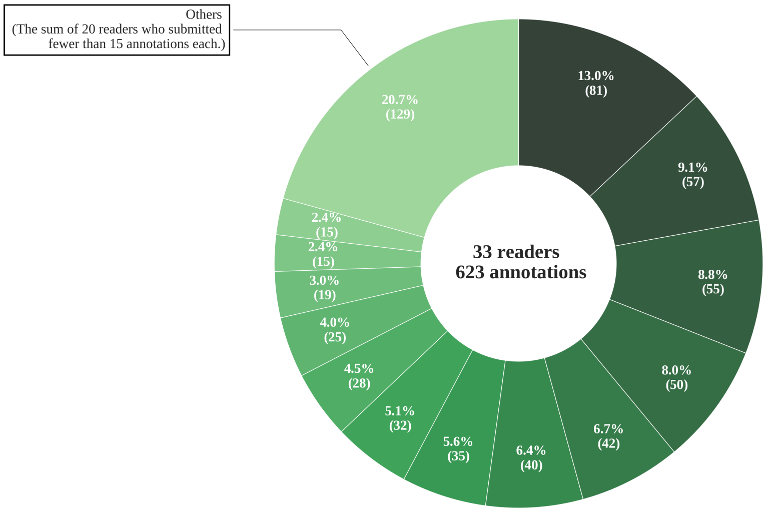


Figure 9: The distribution of the number of annotations per reader as a pie chart. For the 13 readers who submitted at least 15 annotations, the number of annotations and the corresponding percentage of the dataset are shown. Readers who wrote less than 15 annotations were compiled as "others."

identified. We received responses from nine of the 33 readers. In Table 5, we show the demographics of the readers who responded to this additional survey. Seven of the respondents hold master's degrees. Four of them reported that they have a Master of Business Administration. Highly educated readers were potentially attracted to the task because we used academic terms (such as valence and arousal) in the instruction.

4. Analysis

The dataset of shared-character stories collected in this study consisted of short stories. We labeled parts of the dataset with the interestingness and reviews of the stories, scores on each evaluation aspect, and emotions in each sentence. In this section, we discuss our analyses, focusing on emotions, total scores, and scores on the evaluation aspects.

Item	Question Overview	Answer
Gender	Gender from [Female, Male, Other, Prefer not to answer].	Male (5), Female (4)
Age	Age as of April 1, 2018.	31-35 (3), 36-40 (3), 26-30 (2), 51-55 (1)
Location	Country of residence	US (5), India (3), UK (1)
Language Skills	Native Language	English (6), Tamil (3)
Educational Background	Academic degree as of April 1, 2018	Master (7), Bachelor (2)
Reading Preferences	Favorite genre of stories	⋆
	The number of novels read monthly	2 (3), 5 (2), 3 (2), 1 (2)

⋆ There were various answers in the question asking for readers' favorite genre. We show some examples below. Some volunteers responded with multiple genres.

Fiction, Romantic, Nonfiction, Mystery, Science fiction, Fantasy, Crime, "Science fiction, Fantasy", "Adventure, Biography, Mystery, Crime (detective fiction)"

Table 5: The demographics of workers who cooperated with our story evaluation task. In this additional survey, we received self-reported responses from nine of 33 readers. We arranged the answers in descending order.

Stories do not consist of text alone, but are interpreted in collaboration with readers [28]. In short, the readers play important roles in stories. In this study, we aimed, as much as is possible, to perform a universal evaluation. If we consider individual readers, the interestingness of a story depends on who reads it; a story that is very interesting to one reader may bore another. In this study, we focused on whether the story tended to be favorable to many people; therefore, this study did not examine individual preference. We had multiple volunteers annotate the same story and averaged the emotion, total score, and evaluation aspect score values. In other words, we modeled *average readers*.

4.1. Comparison of Human-Written and Artificially Modified Stories

First, we examined data quality by comparing the scores of the 75 human-written and 25 artificially modified stories. In each of the *human-written*, *ran-*

	Interest	Storyness	Fluency	Consistency	Clarity	Meaning	#
Human-written	3.82±1.17	3.85±1.21	4.01±1.06	3.98±1.09	4.00±1.11	3.89±1.08	470
Random last sentence	2.85±1.29	3.28±1.34	3.35±1.23	2.57±1.35	3.10±1.34	2.82±1.23	60
Random last sentence from the same setting	3.48±1.36	3.59±1.35	3.55±1.21	3.48±1.30	3.42±1.28	3.55±1.22	64
Random order	3.41±1.30	3.72±1.33	3.51±1.43	3.62±1.42	3.59±1.32	3.62±1.21	29
All 100 stories	3.68±1.24	3.77±1.25	3.88±1.13	3.78±1.23	3.84±1.20	3.74±1.16	623

Table 6: The results of evaluating human-written and artificially modified (*random last sentence*, *random last sentence from the same setting*, and *random order* condition) stories. The number shown in each cell is the mean ± σ and # indicates the number of submissions in the annotation task.

dom last sentence, *random last sentence from the same setting*, and *random order* conditions, we took the average of the total score (i.e., story interestingness) and scores on the individual evaluation aspects. The results are shown in Table 6.

Human-written stories scored the highest in each aspect and in the total scoring. Stories in the *random last sentence* condition had low *Consistency* and *Meaning* scores: when the last sentence was taken from a different story, readers did not find the stories consistent or meaningful. By contrast, stories in the *random last sentence from the same setting* condition rated higher than those in the *random last sentence* condition in all aspects, possibly because using another story that shares character settings when changing the last sentence minimizes the deviation from the original story as compared to using sentences from other randomly chosen stories. Stories in the *random order* condition were rated higher than those in the *random last sentence from the same setting* condition in terms of *Storyness* and *Consistency*. Based on this result, readers were able to rearrange short stories on their own and read the story in terms of consistency and storyness.

To investigate the answers in more detail, we show the distribution of total

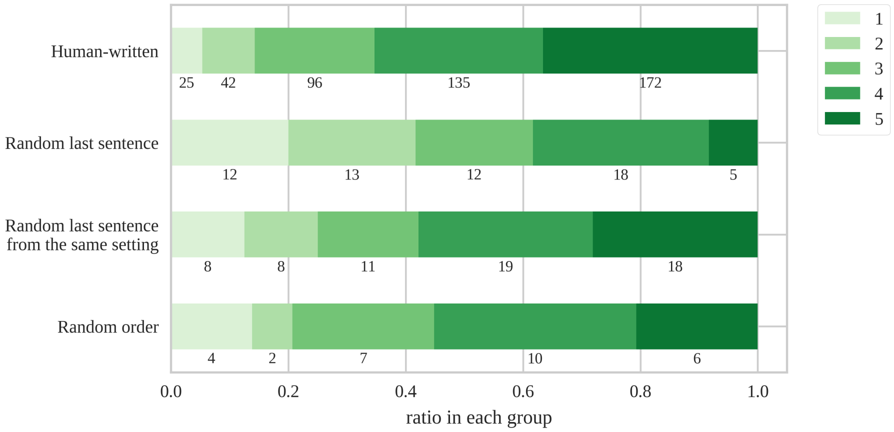


Figure 10: The distribution of the total score (i.e., story interestingness) in each condition. The scores were annotated on a five-point scale (ranging from 1 to 5, the higher the better). The values under the bars are the number of answers.

scores in each condition in Figure 10. As discussed above, *human-written* stories tended to have higher scores. The *random last sentence* condition received lower total scores at a high rate, which is likely the reason for the low average point of this condition in Table 6. It is interesting that even artificially modified stories got scores of "4" at a high rate, perhaps because some MTurk volunteers tended to attach a slightly better score than the central value (in this case, "3") when they were asked to make annotations.

To examine the difference of evaluation among the volunteer writers, we group the *human-written* stories for each writer and show their total scores in Figure 11. In terms of the 75 stories in this condition, 24 writers were included. Although the number of stories was small for each writer, it can be said that there were some writers receiving relatively high scores. We leave it as a future work about how to choose particularly good writers when using crowdsourcing.

An example of a highly rated story is shown in Table 1; its total score was 4.57. The emotional flow of this story is shown in Figure 2.

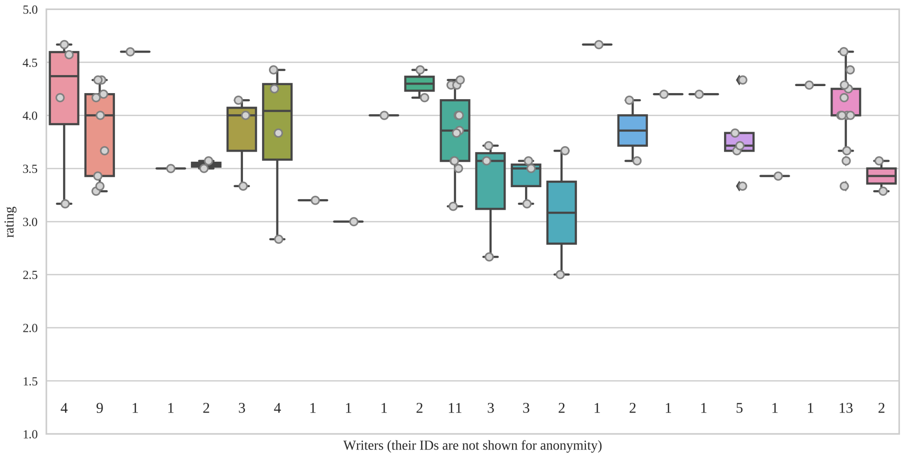


Figure 11: The distribution of the total score in *human-written* stories. There were 75 stories in this condition and 24 writers were included. We group the stories for each writer and show it in a box and whisker plot. For clarity, we add points with jitter on the plot. Numbers shown at the bottom indicate the number of stories each writer wrote. For anonymity, workers' IDs are deleted.

4.2. Regression from Emotions to Total Scores

To better understand the relationship between emotions and the stories' evaluation scores, we entered the stories' emotional tenor into regression models to predict their interestingness. The average scores and values are taken across volunteers for each story. We used random forest (RF) and automatic relevance determination (ARD) regression models with the following parameters: in the RF, the number of decision trees was set to 10; in ARD, the parameters were set to $\alpha_1 = 1e-6$, $\alpha_2 = 1e-6$, $\lambda_1 = 1e-6$, and $\lambda_2 = 1e-6$. Mean squared error (MSE) was used to evaluate these models.

We used the leave-one-out method to split stories into training and test data (99 stories for training and one story for testing). As we used 100 stories for this experiment, there were 100 ways to divide the dataset; we showed the average value of 100 * N divisions. Since RFs have high execution randomness, we set N = 10. For ARD (10), N = 10 to match the number of RF executions, and

		Character 1			Character 2			Reader		
		V	A	V + A	V	A	V + A	V	A	V + A
RF	mean	0.343	0.291	0.301	0.345	0.344	0.322	0.351	0.293	0.262
	SD	0.525	0.415	0.445	0.483	0.498	0.459	0.512	0.426	0.380
	p-value	-	0.000195	0.000177	-	0.915	0.0253	-	3.39e-05	5.67e-17
ARD (10)	mean	0.298	0.260	0.274	0.298	0.273	0.252	0.293	0.232	0.235
	SD	0.424	0.384	0.382	0.423	0.418	0.404	0.427	0.346	0.331
	p-value	-	6.67e-08	0.00227	-	0.000368	8.15e-08	-	4.15e-12	5.67e-09
ARD (1)	mean	0.298	0.260	0.274	0.298	0.273	0.252	0.293	0.232	0.235
	SD	0.424	0.384	0.382	0.423	0.418	0.404	0.427	0.346	0.331
	p-value	-	0.0900	0.338	-	0.263	0.0921	-	0.0295	0.0673

Table 7: The result of regression from emotions to total scores. V indicates using valence, A indicates using arousal, and V + A indicates using both valence and arousal. To show the significance of the difference, we conducted paired t-tests between V and A, between V and V + A, respectively. For p-value, column A shows the result of comparing V and A. The result of comparing V and V + A is shown in column V + A. P-values less than 0.05 are underlined for clarity.

ARD (1) is executed with N = 1.

The results are shown in Table 7. Both models, from any point of view (Character 1, Character 2, or Reader), performed better when considering arousal rather than only valence.

We conducted paired t-tests to investigate the significance of the difference in the regression scores on whether we consider arousal or not. We show p-values in Table 7, where $p<0.05$ are underlined for clarity. In RF, better results were obtained when considering arousal in addition to valence in all cases, except for using only the arousal of Character 2. Although it cannot be said to be significant in ARD (1), we got better result including arousal than using valence alone in ARD (10). In particular, focusing on the reader, the case where only arousal was used was significantly better than the case where only valence was used. From this experimental result, we posit that there is a relationship between story interestingness and readers' emotions. Moreover, emotions should not be limited to only one axis (positive/negative) for story evaluation.

Title: let the best man win

Story: john and David went to the same high school. they were a part of the football team. when the current captain of the team graduated, they were the top contenders. Their friendship got strained because of competition with each other. They decided that the best man would win

Settings: (1) John is a high school student. (2) David is a friend of John.

Evaluation Score (Ranging from 1 to 5: 1 is bad, 5 is good):

Total Score: 4.00, Storyness: 4.00, Frequency: 4.33, Consistency: 4.50, Clarity: 4.50, Meaning: 4.50

Reviews:

- This is a logical story with a clear-cut theme. Two friends briefly become enemies, but resolve their differences through sportsmanship.

- The story is not at all consistent

- not interesting,neutral story

- i love the flow of the story

- THE story was interest ,i feel very happy to winner the match

- This story is about two friends whose friendship was strained by competition, but they decided on a resolution. The way the story illustrates this is good.

Table 8: An example story in which the emotional flow of the characters and readers differed. See Figure 12 for further information.

In what kind of stories do the emotions of the characters and of the reader move in different directions? Table 8 and Figure 12 show an example of a story in which the emotional flow of the characters and that of the readers differed. The fourth sentence states that the characters "friendship was strained." The emotions of the reader went in the leftward and upward (negative, excited) directions. However, in the same sentence, the emotions of Characters 1 and 2 went in the leftward and downward (negative, calm) directions at this point, showing that the emotions of the characters and of the reader are not always synchronized when we focus on arousal. The average value of the total score of this story is 4.0, which means that it was highly evaluated by the volunteers.

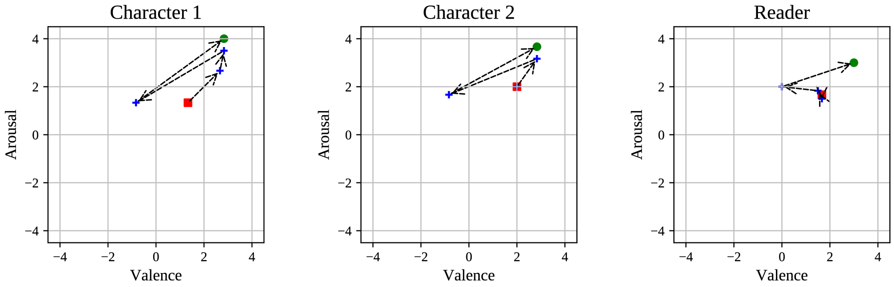


Figure 12: The emotional flow of the story in Table 8, wherein the characters' and readers' emotional flows differed. The emotional flow is illustrated as in Figure 1. The relationship between the two characters becomes strained in the story's fourth sentence and the readers' emotions move leftward and upward (more negative and more excited). However, in the same sentence, the emotions of Characters 1 and 2 move leftward and downward (becoming more negative and calmer).

5. Conclusion

This study makes two main contributions. We proposed and constructed a new dataset with shared-character settings and clarified the importance of emotions when evaluating story interestingness.

First, we proposed a novel approach to story collection, the shared-character story, wherein writers (volunteers working on Amazon MTurk) were given character settings as story prompts. We collected five-sentence-long stories to create our crowdsourced shared-character story dataset. Second, using part of the collected data, we asked volunteers to annotate the stories with evaluation scores, including those measuring emotions and overall characteristics, such as interestingness. Based on the annotated scores and emotions, we performed empirical analyses to examine the relationship between emotions and story interestingness.

Our analyses of the short shared-character stories revealed that the characters' emotions do not necessarily match those of the reader. We also examined the finding that the readers' emotions—rather than the character's emotions—strongly correlate with the story's evaluation score. Here, note that the character settings in shared-character stories are not exhaustive. It is possible that biases with respect to how the settings were developed and the small dataset may have influenced the study's results. Although such restrictions and challenges remain, our shared-character stories can be considered a pioneering dataset for new research topics on the relationship between story evaluations and emotions.

Writing stories is a nearly unconstrained activity. Preparing exhaustive character settings and collecting all possible storylines is not realistic. Therefore, universal methods for creating character settings from which writers can imagine interesting and diverse stories must be considered by generalizing our findings. In future work, we plan to construct a larger dataset using the "shared-character story" method proposed in this paper. With this dataset, we intend to acquire a deeper understanding of the relationship between story evaluations and emotions, because we believe that our findings will aid further research in this direction. As the reader's emotions are correlated with the story's interestingness, estimating a reader's emotion from the text will be an important step for the automatic evaluation of a story.

6. Acknowledgements

This work was partially funded by the ImPACT Program of the Council for Science, Technology, and Innovation (Cabinet Office, Government of Japan) and was partially supported by CREST, JST.

References

[1] J. Campbell, The Hero with a Thousand Faces, Pantheon Books, 1949.

[2] V. I. Propp, Morphology of the Folktale (Translated by L. Scott), University of Texas Press, 1968.

[3] S. Field, The Screenwriter's Workbook, Revised Edition, Delta Trade Paperbacks, 2006.

[4] B. Snyder, SAVE THE CAT! The Last Book on Screenwriting You'll Ever Need, Michael Wiese Productions, 2005.

[5] V. Fortuin, R. Weber, S. Schriber, D. Wotruba, M. Gross, InspireMe: Learning sequence models for stories, in: 30th Conference on Innovative Applications of Artificial Intelligence, 2018.
URL https://www.disneyresearch.com/publication/inspireme/

[6] M. Kapadia, J. Falk, F. Zünd, M. Marti, R. W. Sumner, M. Gross, Computer-assisted authoring of interactive narratives, in: Proceedings of the 19th Symposium on Interactive 3D Graphics and Games, ACM, New York, NY, USA, 2015, pp. 85–92. doi:10.1145/2699276.2699279.
URL http://doi.acm.org/10.1145/2699276.2699279

[7] R. Koncel-Kedziorski, I. Konstas, L. Zettlemoyer, H. Hajishirzi, A theme-rewriting approach for generating algebra word problems, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1617–1628.
URL https://aclweb.org/anthology/D16-1168

[8] C. O. Alm, D. Roth, R. Sproat, Emotions from text: Machine learning for text-based emotion prediction, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005.
URL http://www.aclweb.org/anthology/H05-1073

[9] S. Chaturvedi, H. Peng, D. Roth, Story comprehension for predicting what happens next, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1603–1614.
URL https://www.aclweb.org/anthology/D17-1168

[10] J. Eisenberg, M. Finlayson, A simpler and more generalizable story detector using verb and character features, in: Proceedings of the 2017 Conference

on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2708–2715.
URL https://www.aclweb.org/anthology/D17-1287

[11] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, J. Allen, A corpus and cloze evaluation for deeper understanding of commonsense stories, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 839–849.
URL http://www.aclweb.org/anthology/N16-1098

[12] T. Wang, P. Chen, B. Li, Predicting the quality of short narratives from social media, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI, Melbourne, Australia, 2017, pp. 3859–3865. doi:10.24963/ijcai.2017/539.
URL https://doi.org/10.24963/ijcai.2017/539

[13] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, P. S. Dodds, The emotional arcs of stories are dominated by six basic shapes, EPJ Data Science 5 (1) (2016) 31. doi:10.1140/epjds/s13688-016-0093-1.
URL https://doi.org/10.1140/epjds/s13688-016-0093-1

[14] J. A. Russell, A circumplex model of affect, Journal of personality and social psychology 39 (1980) 1161–1178.

[15] O. Bajgar, R. Kadlec, J. Kleindienst, Embracing data abundance: Booktest dataset for reading comprehension, Computing Research Repository arXiv:1610.00956.

[16] N. Chambers, D. Jurafsky, Unsupervised learning of narrative event chains, in: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 789–797.
URL http://www.aclweb.org/anthology/P/P08/P08-1090

[17] F. Hill, A. Bordes, S. Chopra, J. Weston, The goldilocks principle: Reading children's books with explicit memory representations, Computing Research Repository arXiv/1511.02301.

[18] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, E. Grefenstette, The NarrativeQA reading comprehension challenge, Transactions of the Association for Computational Linguistics.

[19] D. Paperno, G. Kruszewski, A. Lazaridou, N. Q. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, R. Fernandez, The LAMBADA dataset: Word prediction requiring a broad discourse context, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1525–1534.
URL `http://www.aclweb.org/anthology/P16-1144`

[20] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, J. Allen, Lsdsem 2017 shared task: The story cloze test, in: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 46–51.
URL `http://aclweb.org/anthology/W17-0906`

[21] C. Strapparava, R. Mihalcea, Learning to identify emotions in text, in: Proceedings of the 2008 ACM Symposium on Applied Computing, SAC '08, ACM, New York, NY, USA, 2008, pp. 1556–1560. `doi:10.1145/1363686.1364052`.
URL `http://doi.acm.org/10.1145/1363686.1364052`

[22] M. Abdul-Mageed, L. Ungar, Emonet: Fine-grained emotion detection with gated recurrent neural networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada,

2017, pp. 718–728.
URL http://aclweb.org/anthology/P17-1067

[23] P. Ekman, Facial expression and emotion, American Psychologist 48(4) (1993) 384–392.

[24] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, 1st Edition, O'Reilly Media, Inc., 2009.

[25] M. M. Bradley, P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, Journal of Behavior Therapy and Experimental Psychiatry 25 (1) (1994) 49 – 59. doi:https://doi.org/10.1016/0005-7916(94)90063-9.
URL http://www.sciencedirect.com/science/article/pii/0005791694900639

[26] A. S. Cowen, D. Keltner, Self-report captures 27 distinct categories of emotion bridged by continuous gradients, Proceedings of the National Academy of Sciences 114 (38) (2017) E7900–E7909. arXiv:http://www.pnas.org/content/114/38/E7900.full.pdf, doi:10.1073/pnas.1702247114.
URL http://www.pnas.org/content/114/38/E7900

[27] S. Srivastava, S. Chaturvedi, T. Mitchell, Inferring interpersonal relations in narrative summaries (2016).
URL https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12173

[28] U. Eco, Lector in fabula, Gruppo Editoriale Febbri, 1979.