PLOTMACHINES: Outline-Conditioned Generation with Dynamic Plot State Tracking

Hannah Rashkin[1], **Asli Celikyilmaz**[2], **Yejin Choi**[1,3], **Jianfeng Gao**[2]
[1] Paul G. Allen School of Computer Science & Engineering, University of Washington
[2] Microsoft Research, Redmond, WA, USA
[3] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{hrashkin,yejin}@cs.washington.edu, {aslicel,jfgao}@microsoft.com

Abstract

We propose the task of *outline-conditioned story generation*: given an outline as a set of phrases that describe key characters and events to appear in a story, the task is to generate a coherent narrative that is consistent with the provided outline. This task is challenging as the input only provides a rough sketch of the plot, and thus, models need to generate a story by interweaving the key points provided in the outline. This requires the model to keep track of the dynamic states of the latent plot, conditioning on the input outline while generating the full story. We present PLOTMACHINES, a neural narrative model that learns to transform an outline into a coherent story by tracking the dynamic plot states. In addition, we enrich PLOTMACHINES with high-level discourse structure so that the model can learn different writing styles corresponding to different parts of the narrative. Comprehensive experiments over three fiction and non-fiction datasets demonstrate that large-scale language models, such as GPT-2 and GROVER, despite their impressive generation performance, are not sufficient in generating coherent narratives for the given outline, and dynamic plot state tracking is important for composing narratives with tighter, more consistent plots.

1 Introduction

Composing a story requires a complex planning process. First, the writer starts with a rough sketch of what key characters and events the story will contain. Then, as they unfold the story, the writer must keep track of the elaborate plot that weaves together the characters and events in a coherent and consistent narrative.

We study this complex storytelling process by formulating it as the task of *outline-conditioned story generation*, illustrated in Figure 1. Given an outline, a set of phrases describing key characters and events to appear in a story, the task is to generate a coherent narrative that is consistent with the provided outline. This task is challenging as the input provides only the rough elements of the plot[1]. Thus, the model needs to flesh out how these plot elements will intertwine with each other across different parts of the story. The flowchart in Figure 1 demonstrates an example of a latent plot structure: different key phrases from the outline appear and re-appear jointly throughout different sentences and paragraphs. Notably, the way that outline points are interwoven needs to be determined dynamically based on what's already been composed while also staying true to the original outline and overall narrative structure.

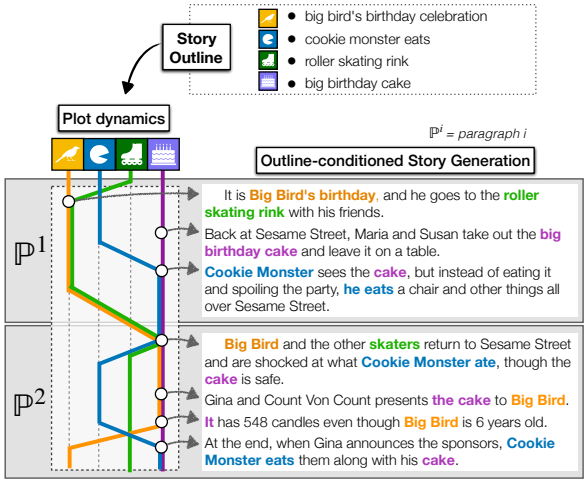

Figure 1: An outline (input) paired with a story (output) from the Wikiplots training set. Plot elements from the outline can appear and reappear non-linearly throughout the plot, as shown in plot dynamics graph. Composing stories from an outline requires keeping track of how outline phrases have been used while writing.

We present PLOTMACHINES, a novel narrative transformer that simulates the outline-conditioned

[1]Here, we define plot as the main sequence of events in the story.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4274–4295,
November 16–20, 2020.

generation process described above.[2] Our model learns to transform an outline into a multi-paragraph story using dynamic memory blocks that keep track of the implicit plot states computed using the outline and the story generated thus far. We draw inspiration from prior work in dialogue state tracking (Thomson and Young, 2010; Lee, 2013; Chao and Lane, 2019), entity tracking (Henaff et al., 2017; Bosselut et al., 2018), and memory networks (Sukhbaatar et al., 2015) for keeping track of plot states. We also inform our model with high-level narrative structure using discourse labels so that it can learn different styles of writing corresponding to different parts of the narrative (i.e. beginning, middle, and end). PLOTMACHINES is, to the best of our knowledge, the first model designed to generate multi-paragraph stories conditioned on outlines and can be trained end-to-end to learn the latent plot patterns without explicit plot annotations for supervision.

To support research on outline-conditioned generation, we present three datasets, including both fiction and non-fiction domains, where multi-paragraph narratives from existing datasets are paired with automatically constructed outlines using state-of-the-art key phrase extraction. Importantly, our task formulation of outline-conditioned generation is general and can be applied to various forms of grounded language generation. Comprehensive experiments on these datasets demonstrate that recently introduced state-of-the-art large-scale language models such as GPT-2 and GROVER (Radford et al., 2019; Zellers et al., 2019), despite their impressive generation performance, still struggle to generate coherent narratives that are consistent with input outlines. Our experiments indicate that dynamic plot state tracking is important for constructing narratives with tighter and more consistent plots compared to competitive baselines.

Our main contributions are: (1) a new task formulation of outline-conditioned story generation, (2) the presentation of three new datasets for this task, (3) PLOTMACHINES, a novel narrative transformer that learns to transform outlines to full stories with dynamic plot state tracking, and (4) empirical results demonstrating the limitations of state-of-the-art large-scale language models and the advantage of PLOTMACHINES compared to competitive baselines.

[2]code available at https://github.com/hrashkin/plotmachines

2 Outline-Conditioned Generation

The Task: Our primary goal is to design a task for investigating how story generation models can plan long narrative according to controllable story elements. To that end, we introduce the outline-conditioned story generation task, which takes a plot outline as input and produces a long, multi-paragraph story.

In order to be flexible to multiple forms of control that might be required for different downstream tasks, we envision plot outlines to be defined loosely as lists of an arbitrary number of un-ordered plot points that should guide a story being generated. Plot points could consist of high-level concepts, low-level events, or even detailed sentences. For practical reasons, in this work, we limit the scope of plot points to events and phrases since these can be automatically extracted. Future work could explore alternate methods of defining plot outlines, perhaps using an event-based planning systems (Porteous and Cavazza, 2009; Riedl, 2009; Riedl and Young, 2010; Fan et al., 2019) for generating key points.

More concretely, in this paper, we formulate the outline as a list of un-ordered bullet points which reflect key phrases to be loosely integrated in the output narrative. These plot outlines are inspired, in part, by previous work in short-form story generation tasks that conditioned on storylines (Peng et al., 2018; Yao et al., 2019), which were defined as an ordered list of exactly five single-word points. We extend this concept to long-form story generation by defining a plot outline more flexibly as: an *un-ordered* list of *an arbitrary number of multi-word* plot elements. An outline also differs from a writing prompt, such as those found in other controllable writing tasks (Fan et al., 2018), which are more abstract and often just a starting point for a story. Unlike a prompt, an outline is a list of concrete points that must appear somewhere in the narrative.

One challenge of this task is to create stories that have appropriate discourse and narrative flow. A second challenge is for stories to include the outline in a natural way. For example, it may be appropriate for certain outline points to be used only later on in the story (e.g. the protagonist dying may be more typically used at the end).

Wikiplots # stories : 130k avg # pars : 3.1 data-split : 90/5/5	**Outline**: • the rocky horror picture show • convention attendees includes servants (...) **Story:** A criminologist narrates the tale of the newly engaged couple, Brad Majors and Janet Weiss, who find themselves lost and with a flat tire on a cold and rainy late November evening, somewhere near Denton in 1974 (...)
WritingPrompts # stories : 300k avg # pars : 5.9 data-split : 90/5/5	**Outline**: • found something protruding • geometric shapes glowing • sister kneeling beside • dead bodies everywhere • darkness overwhelmed • firelight flickering (...) **Story**: It was dark and Levi was pretty sure he was lying on his back . There was firelight flickering off of what was left of a ceiling . He could hear something but it was muffled . He (...)
NYTimes # stories : 240k avg # pars : 15.2 data-split : 90/5/5	**Outline**: • upcoming annual economic summit meeting • take intermediate steps (...) **Article:** The long-simmering tensions in Serbia's province of Kosovo turned violent in recent weeks and threaten to ignite a wider war in the Balkans. Only a concerted diplomatic effort by the United States can keep the conflict from escalating. Though he has been attentive to the problem (...)

Table 1: Datasets used in the experiments showing the number of stories, the average number of paragraphs per story, and the split of stories across train/dev/test. We also show an example outline and a short excerpt from a story. We show examples of the full stories in the Supplementary Material.

Dataset: Outline to Story: We construct three datasets for outline-conditioned generation[3] by creating novel plot outlines to be used as inputs to generating stories from three existing story datasets. Table 1 shows statistics and examples from each dataset. We focus on fictitious generation, but also include the news domain for generalization. We build on existing story datasets for the target narratives, which we pair with automatically constructed input outlines as described below:

Wikiplots corpus[4] consists of plots of TV shows, movies, and books scraped from Wikipedia.

WritingPrompts (Fan et al., 2018) is a story generation dataset, collected from the /r/WritingPrompts subreddit – a forum where Reddit users compose short stories inspired by other users prompts. We use the same train/dev/test split from the original dataset paper.

NYTimes (Sandhaus, 2008) contains news articles rather than fictional stories, unlike the other two datasets.[5]

Outline Extraction We extract a list of plot outlines from each dataset to use as input using the RAKE (Rapid Automatic Keyword Extraction) algorithm (Rose et al., 2010)[6]. RAKE is a domain independent keyword extraction algorithm, which determines key phrases in a document based on the word frequency and co-occurrence statistics. We filtered key-points with overlapping n-grams. This is inspired by similar RAKE-based methods for creating storylines (Peng et al., 2018), but differs in that we extract longer outline points (3-8 words each) with no particular order.

3 PLOTMACHINES

Our approach to this task is to design a model that combines recent success in text generation with transformer-based architectures (Vaswani et al., 2017) with memory mechanisms that keep track of the plot elements from the outline as they are used in the story. We also incorporate special discourse features into the modelling to learn a structure over the long multi-paragraph story format.

We introduce PLOTMACHINES (PM), an end-to-end trainable transformer built on top of the GPT model[7] (Radford et al., 2018), as shown in Figure 2. Given an outline as input, the model generates paragraphs, recurrently, while updating a memory matrix M that keeps track of plot elements from the outline. This generation framework is motivated by human writing styles, in which each paragraph is a distinct section of related sentences.

At each time step, i, PLOTMACHINES generates a new paragraph $\mathbb{P}^i$:

$$(\mathbb{P}^i, h^i, M^i) = \text{PM}(o, d^i, h^{i-1}, M^{i-1}) \quad (1)$$

where o is the outline representation (Sec. 3.1), d^i is the discourse representation associated with paragraph i (Sec. 3.2), h^{i-1} is a vector representation of the preceding story context (Sec. 3.3), and M^{i-1} is the previous memory (Sec. 3.4).

3.1 Outline Representation

The plot outline (i.e. the input to the model) is treated as a sequence of tokens, o, and used as input for the transformer for each paragraph that is

[3]Code for replicating data creation available at www.github.com/hrashkin/plotmachines

[4] www.github.com/markriedl/WikiPlots

[5]Due to concerns over fake news creation, we will not release the model trained on this data.

[6]https://pypi.org/project/rake-nltk/

[7]We build on top of GPT, though our approach could be used with most transformer-based LMs. In experiments, we also look at a version of PLOTMACHINES using GPT-2 (Radford et al., 2019) as a base.

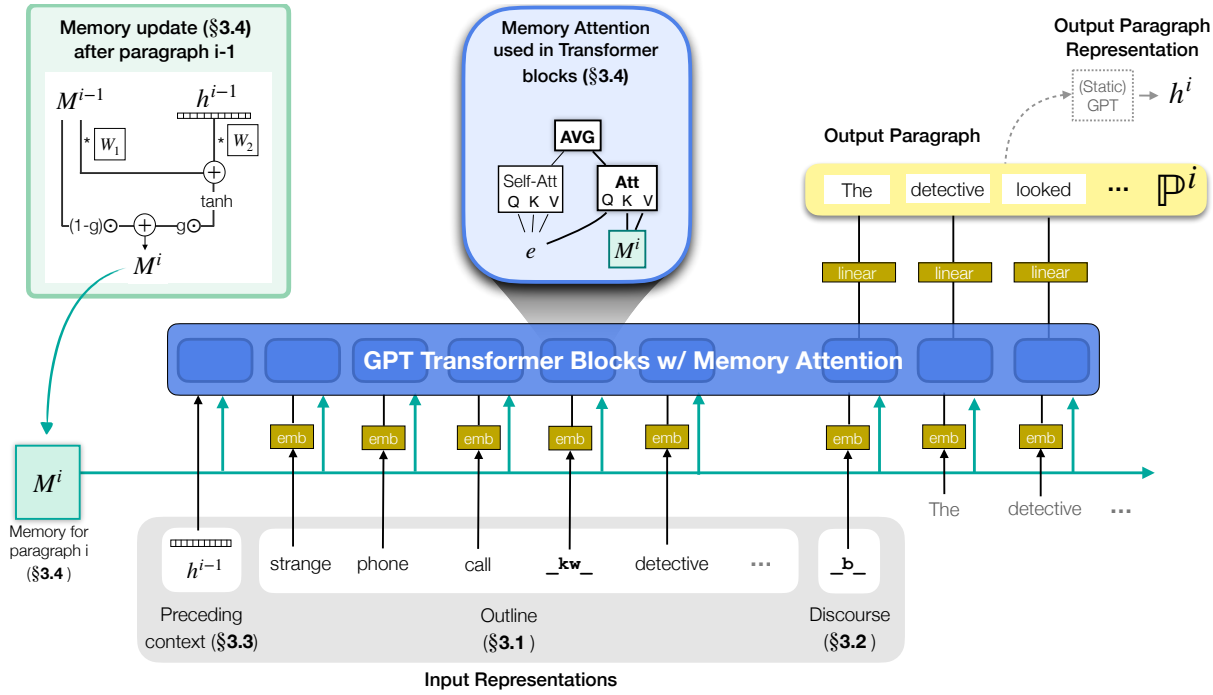

Figure 2: PLOTMACHINES: The model generates a paragraph $\mathbb{P}^i$ using the memory (M^{i-1}), the previous paragraph representation (h^{i-1}), the outline representation (o) and discourse representation (d^i). First, a gated update mechanism updates the memory using the previous memory and previous paragraph representation. Each transformer block includes extra attention over the current memory matrix M^i. The previous paragraph representation, h^{i-1}, the outline, and discourse tag (e.g. _b_) are also prepended to the generation as an input sequence (gray box). The output tokens of the generated paragraph are used to compute h^i using a static GPT model.

generated. We use special **_kw_** tokens to delimit each plot point in the outline and end the sequence with a special **_endkw_** token. We truncate the entire outline to maximum of n tokens. For example, an outline containing two plot points ({'strange phone call', 'detective'}) is turned into the input sequence:

`strange phone call _kw_ detective _endkw_`

3.2 Discourse Representation

We posit that there are stylistic differences in how the beginning, middle and end of a story are written. To learn these differences, we introduce d^i, discourse information about whether the i-th paragraph is an introduction, body, or conclusion paragraph. We append a special token to the outline representation as part of the input sequence: **_i_**, **_b_**, **_c_** for the introduction, body, and conclusion paragraphs respectively[8].

3.3 Preceding Context Representation

With the goal of incorporating previous story context in generating each paragraph, we use h^{i-1}, an embedded representation of the previous paragraph, which is added to the model input. More concretely, h^{i-1} is computed as the average embedding of GPT output representations of words from the previous paragraph (using a GPT model that is static, i.e. not finetuned). The h^{i-1} vector is used as an initial input to the transformer architecture, as shown in Figure 2.

3.4 Memory Representation

We implement memory to address two key task challenges. First, we want to keep track of the portions of the outline that have been mentioned. Second, we want to maintain semantic coherence across the entire story. To address these two challenges, we implement the memory as consisting of two parts: K, a set of vectors keeping track of outline points, and D, a matrix that stores a latent topic distribution of what's been written so far.

Notation: We define d as the embedding size of the transformer model and n as the maximum number of tokens in the outline. Memory is treated as a $\mathbb{R}^{d \times 2n}$ matrix which consists of two smaller matrices stacked together ($M = [K; D]$). K is a $\mathbb{R}^{d \times n}$ representation of outline points and D is a $\mathbb{R}^{d \times n}$ representation of the latent document state. K is

[8] We make the simplifying assumption that the first paragraph is an introduction, the last paragraph is the conclusion paragraph, and the other paragraphs are all body paragraphs.

initialized with embeddings representing each of the tokens in the outline and D is randomly initialized. The j-th column of memory at the timestep for paragraph i will be denoted M_j^i.

Updating memory: The memory is updated (top left corner of Fig. 2) using h^{i-1}, the average GPT output representation of the previous paragraph. We use update equations based on those in entity-based models such as Henaff et al. (2017). We use a gating mechanism, g, to allow the model to learn to flexibly control how much each cell in memory is updated, as below:

$$\hat{M}_j^i = tanh(W_1 M_j^{i-1} + W_2 h^{i-1}) \quad (2)$$

$$g_j^i = sigm(W_3 M_j^{i-1} + W_4 h^{i-1}) \quad (3)$$

$$M_j^i = (1 - g_j^i) \odot M_j^{i-1} + g_j^i \odot \hat{M}_j^i \quad (4)$$

where all W's are matrices of dimension $\mathbb{R}^{d \times d}$.

Transformer Blocks with Memory: Lastly, we must alter the GPT transformer blocks to include the memory in the language modeling. We alter the attention used within the transformer blocks to contain two parallel attention modules, as shown in Figure 2. One attention module (on the left in the figure) performs the standard GPT self-attention using transformer inputs to create queries, keys, and values. The other attention module uses transformer input to attend over the memory vectors (i.e., using the memory for creating key and value vectors). The outputs of both attention modules are averaged[9] before performing the remaining transformer block operations.

3.5 Training and Decoding

At training time, the model is trained end-to-end on the cross-entropy loss of predicting each paragraph. Gold representations of previous paragraphs in the story are used to update the memory and compute h^{i-1}. At decoding time, the model must decode a document starting with the first paragraph and use its own predictions to compute h^{i-1} and update the memory. Additionally, at decoding time, we assume a five paragraph structure (introduction, three body paragraphs, and conclusion) as a pre-set discourse structure to decode from.

4 Experiments

We present experiments comparing PLOTMACHINES with competitive baselines and ablations using automatic metrics and human judgements targeting multiple aspects of performance. In Sec. A.3, we also include example generations.

4.1 Experimental Set-up

Baselines: We compare with two models that have been used in related conditional story generation tasks. First, we train a Fusion model, from the original WritingPrompts dataset paper (Fan et al., 2018), using delimited outlines as a single input in place of a prompt. We also compare with the static storyline-to-story variant of Plan-and-Write (P&W-Static) from Yao et al. (2019), which is an LSTM-based model that we train by using the plot outline as delimited input.

Additionally, given the recent successes in text generation using large pre-trained LM's, we compare with these models, as well. We finetune the large-scale GROVER (Zellers et al., 2019) (equivalent to GPT-2 medium, 345M param) , which is a transformer-based language model that has been pre-trained for controllable text generation. To finetune GROVER, we give the outline as a delimited form of metadata. GROVER (345M param) has significantly larger capacity than PLOTMACHINES (160M param). Therefore, for more direct comparison, we also investigate a 460M parameter version of PLOTMACHINES that is built on top of GPT-2 medium (Radford et al., 2019) instead of GPT.

Unlike our models, the baselines are trained with the traditional generation framework, to generate an entire document conditioned on outlines without generating each paragraph recurrently.

Ablated PLOTMACHINES Models: We also show results in Table 2 on ablated versions of our model. First, we use the base GPT and GPT2 models, that are fine-tuned similarly to our model but using only outline inputs (without memory, preceding context, or discourse representations). Second, we investigate the effects of using the preceding context representation but still excluding memory and discourse tokens (**PM-NOMEM-NODISC**). Lastly, we use **PM-NOMEM**, a model variant that excludes the memory but uses outline, discourse, and preceding context representations as input.

Details: We use the HuggingFace implementations of GPT and GPT-2, and we fine-tune using ADAM. For generating with our models, we use nucleus sampling with repetition penalties (Holtzman et al., 2019; Keskar et al., 2019) using $p = 90$ and $\theta = 1.5$ for GPT and $p = 70$ and $\theta = 1.4$ for

[9] We experimented with a few other variants of implementing multiple attention mechanisms within the transformer blocks, but found this to be empirically effective.

	Wikiplots			WritingPrompts			New York Times		
Model	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
P&W-Static (Yao et al., 2019)	17.0	3.3	13.6	19.2	3.6	14.4	19.3	4.6	15.6
Fusion (Fan et al., 2018)	22.7	6.0	17.4	14.3	1.7	9.6	**23.2**	**7.2**	**18.1**
GROVER (Zellers et al., 2019)	19.6	5.9	12.5	23.7	5.3	17.2	20.0	5.8	14.2
PLOTMACHINES (GPT)	20.2	5.3	16.0	30.5	5.3	25.4	21.2	5.0	15.5
– base (GPT) (Radford et al., 2018)	13.2	2.0	7.9	22.1	2.7	14.3	13.9	1.6	8.3
PLOTMACHINES (GPT-2)	**22.8**	**6.5**	**17.5**	**31.1**	**6.7**	**26.1**	22.1	6.4	16.5
– PM-NOMEM (GPT-2)	20.5	4.9	15.5	26.6	3.7	23.5	20.0	5.4	14.4
– PM-NOMEM-NODISC (GPT-2)	19.3	1.7	13.9	26.8	4.5	23.2	18.4	3.4	14.2
– base (GPT-2) (Radford et al., 2019)	18.5	3.9	13.3	26.5	4.6	20.5	19.2	4.7	13.6

Table 2: ROUGE Results on Wiki, WritingPrompts and NYTimes Datasets. The top block represents the baseline models on story/article generation, while the bottom blocks include ablations of our PLOTMACHINES models.

GPT-2 (based on a hyperparameter sweep using grid search with cross-entropy on the dev. data). We use a minimum sequence length of 100 bpe tokens per paragraph and a maximum sequence length of 400, 922 bpe per paragraph for GPT and GPT-2, respectively. We set n, the maximum number of outline tokens and memory dimensions to 100. We used the settings for the baselines from their respective papers and codebases.

4.2 Automatic Metrics

In this section, we evaluate performance using different automatic metrics. We compute ROUGE scores (Lin, 2004) and self-BLEU (Zhu et al., 2018) following from previous work (Shen et al., 2019; Zhu et al., 2018) showing that a large ROUGE score together with a low self-BLEU score can demonstrate a model's ability to generate realistic-looking as well as diverse generations.

Coverage We compute ROUGE scores (Lin, 2004) with respect to the gold stories (Table 2). Results show that the full PLOTMACHINES achieves comparable or higher ROUGE on all three datasets. Both PLOTMACHINES variants (using GPT or GPT-2 as a base) achieve improvements over GROVER, even though GROVER includes significantly more parameters than the model using GPT.

Ablations In the bottom block of Table 2, we compare performance of ablated versions of PLOTMACHINES. First, we compare GPT-2 with PM-NOMEM-NODISC, which differs by including preceding context representations. We observe that PM-NOMEM-NODISC performs slightly better than GPT-2, emphasizing the importance of including context from the previous paragraph. Second, we investigate the impact of discourse structure representations. We compare PM-NOMEM-NODISC, which omits the discourse token, with PM-NOMEM, which uses the discourse token. As shown in Table 2, PM-NOMEM generally has higher ROUGE scores than PM-NOMEM-NODISC, indicating that the discourse representation is beneficial to the model. Lastly, we compare PM-NOMEM with the full PLOTMACHINES to determine the effects of having a memory component. Our full model with memory has large ROUGE score improvements over PM-NOMEM, underscoring the importance of the plot state tracking.

Diversity We evaluate the diversity of generated paragraphs from our models using self-BLEU scores (Zhu et al., 2018). In Table 3, we report the self-BLEU scores along with the average length of each generated story. Using all the generated documents from a model, we take one generated document as hypothesis and the others as reference, and calculate BLEU score for every generated document, and define the average BLEU score to be the self-BLEU of the model. While the Fusion model achieved relatively high ROUGE scores, it has generally worse diversity scores (much higher self-BLEU in Table 3). It may be that this model's high ROUGE scores were obtained by producing text that is more repetitive and generic.[10] In contrast, PLOTMACHINES generally achieves good performance on both ROUGE and diversity scores, with self-BLEU scores that are lower than most other models. Notably, they generally have more similar self-BLEU scores to the actual gold stories, indicating that the language diversity is more similar to what humans write.

4.3 Human Evaluations

Due to the limitations of automatic metrics, we also perform extensive human evaluations. We conduct human studies to explore how generated stories compare along three dimensions: outline

[10] We show an example output from the Fusion model in Figure 13.

	Wikiplots					Writing Prompts					NY Times				
Model	AvgL	B-2	B-3	B-4	B-5	AvgL	B-2	B-3	B-4	B-5	AvgL	B-2	B-3	B-4	B-5
Gold Test	330	.74	.50	.29	.15	661	.82	.61	.40	.25	315	.73	.50	.32	.21
P&W-Static	352	.93	.85	.75	.64	675	.97	.94	.89	.85	352	.93	.85	.74	.63
Fusion	191	.84	.71	.58	.48	197	.93	.85	.75	.65	171	.89	.80	.70	.60
GROVER	835	.72	.49	.48	.37	997	.88	.72	.52	.34	719	.79	.57	.38	.25
GPT	909	.77	.47	.25	.11	799	**.73**	**.40**	**.19**	**.08**	739	**.68**	**.36**	.27	**.08**
GPT-2	910	.60	.26	.10	.03	799	.74	.41	**.19**	**.08**	756	.69	**.36**	**.17**	**.08**
PLOTMACHINES (GPT)	682	.77	.58	.40	.27	850	.89	.81	.72	.63	537	.85	.69	.53	.40
PLOTMACHINES (GPT-2)	553	**.56**	**.19**	**.07**	**.02**	799	.83	.56	.30	.14	455	.79	.57	.37	.23

Table 3: Average length of the generated test documents (**AvgL**) and Self-BLEU n-gram (**B-n**) scores on 1000 generated story samples from the test sets. We also include the average length and self-BLEU scores of the gold test data. A lower self-BLEU score together with a large ROUGE (see Table 2) score can justify the effectiveness of a model. We bold the lowest model score in each column; however, we note that sometimes the model self-bleu scores can be lower than in the gold document.

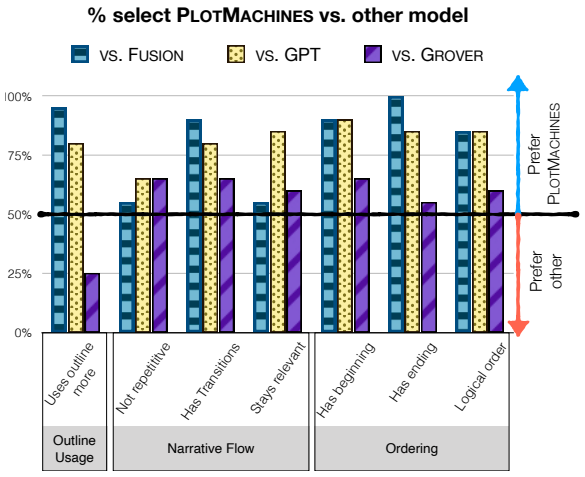

Figure 3: Head-to-head comparsion of PLOTMACHINES vs. three other models for full stories. In the appendix, we report results with standard error metrics (Table 9).

utilization, narrative flow, and ordering. We ask human raters[11] to evaluate each single or pair of generations from the Wikiplots test set, collecting 3 responses per story.[12]

To scale up the crowdsourcing work pragmatically, we split evaluations into two studies: one small-scale study evaluating full-length stories, and one large-scale study evaluating single-paragraph excerpts. In the first study, humans perform head-to-head ratings of 20 randomly sampled stories per pair of models. In the second study, humans rate story excerpts from 100 randomly sampled outputs per model.

[11]using Amazon Mechanical Turk. We include screenshots from the tasks Appendix A.2. In total, over 700 humans participated in all of our studies.

[12]For sampling from Fusion, we restrict to stories or excerpts with two or fewer unk tokens to ensure legibility for workers.

Outline Utilization		
Model A	Model B	% Prefer Model A
Random Paragraph		
PLOTMACHINES	Fusion	**80**% ±4.0
PLOTMACHINES	GPT	**72**% ±4.5
PLOTMACHINES	GROVER	49% ±5.0
Closest Paragraph		
PLOTMACHINES	Fusion	**83**% ±3.8
PLOTMACHINES	GPT	**83**% ±3.8
PLOTMACHINES	GROVER	**54**% ±5.0

Table 4: Humans judge which of two paragraphs better utilize the outlines (when shown either random paragraphs or the paragraphs most similar to the outline).

4.3.1 Full Story Ratings

We give human raters a pair of stories generated from the same outlines and ask them to choose which one is better in different aspects related to outline utilization, narrative flow, and ordering. In Figure 3, we show how often PLOTMACHINES (PM) was selected over the other models (values above 50% indicate that PM was preferred) using the majority vote for each example. PM was selected over base GPT and Fusion in all of the categories, demonstrating that the memory and discourse features are vitally important to improving the base model. While humans rated GROVER as using the outline more, PM is ranked higher in all of the questions about narrative flow and ordering.

4.3.2 Excerpt Ratings

Outline Usage We give raters two paragraphs each generated by different models and ask them to select which is utilizing the outline better. We perform two trials, one with random paragraphs from each story and one with the paragraph from each story that has the most n-gram overlap with the outline (i.e. the closest). In both cases, we compute the majority vote over the three responses and report the percentage of examples where our model is preferred. Results in Table 4 show that, when looking at single paragraphs, humans tend

Model	Narrative Flow			Order
	Rep(↓)	Tran(↑)	Rel(↑)	Acc(↑)
Fusion	2.61	2.98	3.36	**73**
GPT	**1.39**	1.89	2.06	42
GROVER	1.78	3.00	3.29	62
PM	1.64	**3.02**	**3.39**	59

Table 5: Human evaluations of paragraph excerpts from Fusion, GPT, GROVER and PLOTMACHINES (PM) outputs. Narrative flow questions rate the repetitiveness between paragraphs, transitioning, and relevance within paragraphs. In Table 10 of the appendix, we include standard error metrics.

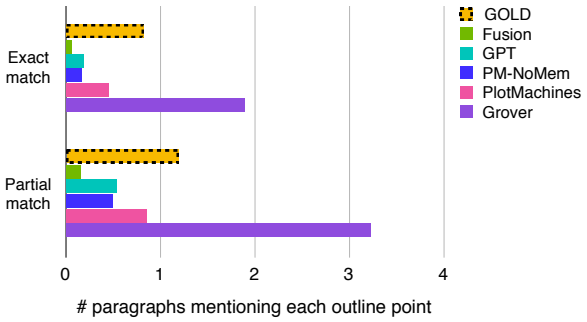

Figure 4: Number of paragraphs mentioning each outline point. PLOTMACHINES with memory covers points more similarly to the gold story, whereas GROVER tends to over-repeat outline points (twice as much as the gold reference).

to choose our PM as using the outlines in a more natural way, particularly when looking at the "closest" paragraph from both models. Fusion and GPT, in particular, are judged to be utilizing the outline much less than PLOTMACHINES.

Narrative Flow In this task, we give raters a generated paragraph (with the previous paragraph as context). They are asked to rate on a scale from 1 to 5 how much the paragraph: (a) repeats content from the previous paragraph, (b) transitions naturally from the previous paragraph, and (c) stays relevant and on-topic throughout the paragraph.

In the left side of Table 5, we show the average ratings of each model. GPT is the least repetitive between paragraphs but has very low subscores for transitions and relevance. We posit that this behavior is likely due to GPT often generating unrelated content from one paragraph to the next. PM tends to have the highest rated transitions and achieve highest relevancy within paragraphs while being much less repetitive between paragraphs than GROVER or Fusion.

Ordering It's challenging for humans to directly rate the ordering of a story based on a short excerpt. We instead set up a related proxy task: we give raters a pair of consecutive generated paragraphs, presented in a random order, and ask them to attempt to decipher the order. The intuition is that if the model output is very well-structured then it should be easier for humans to decipher the order. We compute the accuracy of the majority vote compared to the actual order in the right side of Table 5. Accuracy for PM approaches 60% accuracy and is much better than the base GPT. GROVER and Fusion are easiest for humans to re-order (62%, 73% respectively). This result differs slightly from the full story analysis where the humans preferred PM over GROVER and Fusion in the ordering-based questions. One possible explanation is that these two models, which decode word-by-word, without an explicit notion of paragraph, may be better at resolving coreference problems between paragraphs. This may make it easier for humans to re-order short excerpts even though they generally prefer the overall narrative order of PM due to it having better beginnings, endings, etc. (as indicated in our full story human study).

4.4 N-gram Based Outline Usage Analysis

We perform an additional quantitative study to further investigate how outline points are used in generated stories. For fifty stories in the Wikiplots dev. set, we compute how many paragraphs mention each outline point using exact matching or partial matching ($> 20\%$ of the n-grams in the outline point also appear in the paragraph). We report the results in Figure 4.

We observe that GROVER tends to over-repeat outline points (about twice as much as the gold story). This mirrors our human evaluations that GROVER is more repetitive. This may also explain why human raters in the full story ratings in Sec. 4.3.1 judged GROVER as using the outline more but having worse narrative flow and order. Similar observations have been made about pretrained language models in See et al. (2019) that the models followed story prompts very closely but often copied too much compared to human writing.

In contrast, the Fusion model tends to leave out portions of the outline. This may reflect the way Fusion was originally designed – for use with a task using more abstract prompts as input. The GPT and PM-NOMEM models, while more inclusive than Fusion, are also likely to exclude outline points. The full PM model is generally more inclusive and more similar to the gold reference than the other models. The gold story mentions each outline point

in around one paragraph on average, indicating that there is an ideal balance between the more conservative coverage achieved by our model and the over-repetitive coverage of GROVER.

4.5 Qualitative Examples

In the Appendix, (Sec. A.3), we include examples of model outputs on the validation set with annotations for incorporated outline points. Examples indicate that GROVER often finishes the story and then starts a new story partway through the document. This may help explain why GROVER over-repeats outline points and why humans judge it to be more repetitive and less consistently relevant. In contrast, our model adheres more to a beginning-middle-ending structure.

We also look at examples of introduction and conclusion paragraphs generated by PLOTMACHINES, investigating the discourse the model has learned (Table 11). The model often starts stories by setting the scene (e.g. "In the early 1950s, a nuclear weapons testing continues ....") and tends to write conclusions with a definitive closing action (e.g. "... the film ends with humperdinck and buttercup riding off into the sunset.")

5 Related Work

State Tracking There is a plethora of work in state tracking for dialogue where memory states are updated after each utterance (Thomson and Young, 2010; Young et al., 2010; Lee, 2013; Chao and Lane, 2019). Similarly, SC-LSTMs (Wen et al., 2015) dynamically updated dialogue act representations as a form of sentence planning in spoken dialogue generation. Memory and entity networks (Henaff et al., 2017; Sukhbaatar et al., 2015) and neural checklists (Kiddon et al., 2016) also used similar methods for tracking entities for other tasks. We adapt these techniques for generating stories while tracking plot state that is updated after each paragraph. Our method of decoding paragraphs recurrently also draws on existing work in hierarchical decoding (Li et al., 2015; Shen et al., 2019), which similarly decodes in multiple levels of abstraction over paragraphs, sentences, and words.

Controllable Story Generation There has been a variety of work focusing on generating stories in plot-controllable, plan-driven, or constrained ways (e.g. (Riedl and Young, 2010; Fan et al., 2018; Peng et al., 2018; Jain et al., 2017; Lebowitz, 1987; Ippolito et al., 2019; Pérez y Pérez and Sharples, 2001)). Similar work in creative generation has conditioned on keywords for poetry generation (Yan, 2016; Ghazvininejad et al., 2016; Wang et al., 2016). Outline-conditioned generation is complementary to these tasks in that outlines provide more flexibility than very fine-grained srl-based, event-based, or graph-based plans (Fan et al., 2019; Martin et al., 2017; Harrison et al., 2017; Li et al., 2013) and more structured grounding than coarse-grained prompts (Fan et al., 2018; Xu et al., 2018) or ending goals (Tambwekar et al., 2019). Another similar task generates five line stories from five keywords (Peng et al., 2018; Yao et al., 2019). We generalize to a similar set-up for long-form narratives. Similar to many recent works in this area, we use seq2seq-based approaches, implemented using transformers. We further expand upon the modeling for the challenges specific to our task by using state tracking and applying discourse structure.

6 Conclusion

We present outline-conditioned story generation, a new task for generating stories from outlines representing key plot elements. We facilitate training by altering three datasets to include plot outlines as input for long story generation. In order to keep track of plot elements, we create PLOTMACHINES which generates paragraphs using a high-level discourse structure and a dynamic plot memory keeping track of both the outline and story. Quantitative analysis shows that PLOTMACHINES is effective in composing tighter narratives based on outlines compared to competitive baselines.

Acknowledgements

We would like to thank anonymous reveiwers for their insightful feedback. We also thank Rowan Zellers and Ari Holtzman for their input on finetuning GROVER and other language models, Maarten Sap for his feedback on human evaluations, and Elizabeth Clark for consulting on baselines and related work. We would also like to thank various members of the MSR AI and UW NLP communities who provided feedback on various other aspects of this work. This research was supported in part by DARPA under the CwC program through the ARO (W911NF-15-1-0543), DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031), and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256082.

References

Antoine Bosselut, Over Levy, Ari Holtzman, Coin Ennist, Dieter Fox, and Yejin Choi. 2018. Simulating action dynamics with neural process networks. In *ICLR*.

Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Interspeech*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *ACL*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *ACL*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *EMNLP*.

Brent Harrison, Christopher Purdy, and Mark O. Riedl. 2017. Toward automated story generation with markov chain monte carlo methods and deep neural networks.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *ICLR*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. 2019. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, pages 37–43, Minneapolis, Minnesota. Association for Computational Linguistics.

Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. *ArXiv*, abs/1707.05501.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Chloe Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *EMNLP*.

Michael Lebowitz. 1987. Planning stories. In *Proceedings of the 9th annual conference of the cognitive science society*, pages 234–242.

Sungjin Lee. 2013. Structured discriminative model for dialog state tracking. In *SIGDIAL*.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark O. Riedl. 2013. Story generation with crowdsourced plot graphs. In *AAAI*.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, Beijing, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. https://pypi.org/project/rouge/.

Lara J. Martin, Prithviraj Ammanabrolu, William Hancock, S. Singh, Brent Harrison, and Mark O. Riedl. 2017. Event representations for automated story generation with deep neural nets. In *AAAI*.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*.

Rafael Pérez y Pérez and Mike Sharples. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, 13:119 – 139.

Julie Porteous and Marc Cavazza. 2009. Controlling narrative generation with planning trajectories: the role of constraints. In *Joint International Conference on Interactive Digital Storytelling*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Mark O Riedl. 2009. Story planning: Creativity through exploration, retrieval, and analogical transformation. In *Minds and Machines*, volume 20(4):589614.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. In *Journal of Artificial Intelligence Research*.

Stuart Rose, Nick Cramer, and Dave Engel. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory edited by Michael W. Berry and Jacob Kogan, John Wiley & Sons, Ltd*.

Evan Sandhaus. 2008. The new york times annotated corpus ldc2008t19. *DVD. Philadelphia: Linguistic Data Consortium*.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In *CoNLL*, volume abs/1909.10705.

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. 2019. Towards generating long and coherent text with multi-level latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2079–2089, Florence, Italy. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NeurIPS*, pages 2440–2448.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2019. Controllable neural story plot generation via reward shaping. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5982–5988. International Joint Conferences on Artificial Intelligence Organization.

Blaise Thomson and Steve J Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech and Languag*, 24(4):562588.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. In *Coling*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics.

Rui Yan. 2016. i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In *IJCAI*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *AAAI*, pages 7378–7385.

Steve Young, Milica Gasic, Simon Keizer, Francois Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, 24(2):150174.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *NeurIPS*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *CoRR*, abs/1802.01886. https://github.com/geek-ai/Texygen.

A Supplementary Materials

A.1 Examples from Training Datasets

We show full stories in Tables 6-8 corresponding to the excerpts shown in the Dataset sub-section of Outline-Conditioned Generation in the main text.

A.2 Human Evaluation Details

In Figures 5-8, we show the questionaires we asked the human raters. In question 2 of the full story task, we asked about which story was *more* repetitive, but we flip their answers in Figure 3 to show the model that was *less* repetitive in the Figure (i.e. for ease of reading, we made higher better as with the other metrics).

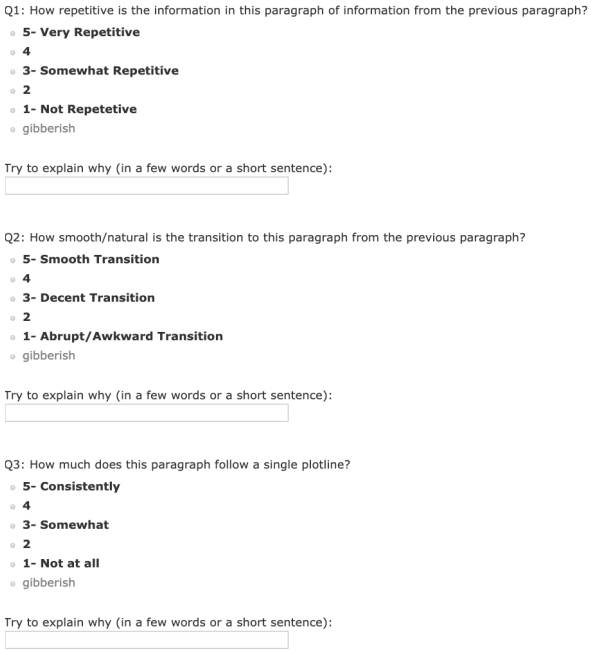

Figure 5: Questionaire for the narrative flow questions about paragraph excerpts. We pay humans $1.00 per HIT.

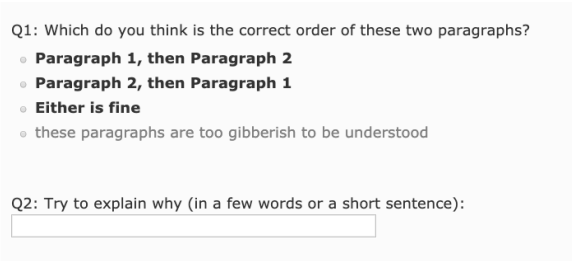

Figure 6: Questionaire for the ordering questions about paragraph excerpts. We pay humans $1.00 per HIT.

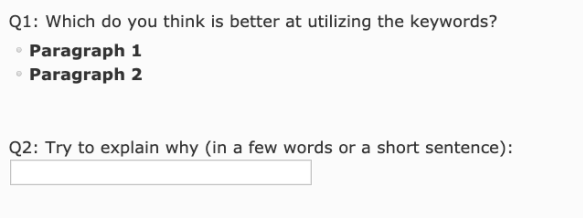

Figure 7: Questionaire for the head-to-head outline usage questions about paragraph excerpts. We pay humans $1.00 per HIT.

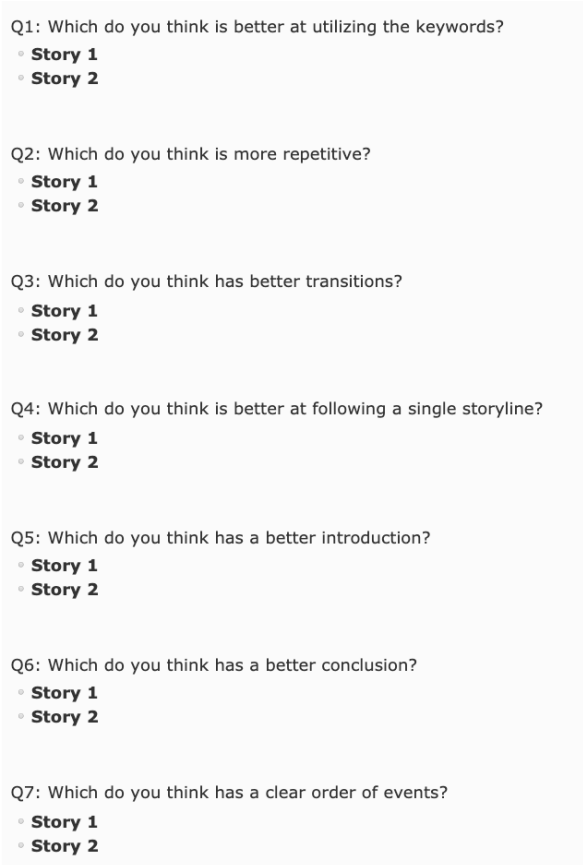

Figure 8: Questionaire for the head-to-head questions about full stories. We pay humans $2.00 per HIT. Note: we reversed the answers to question 2 so that we could show which models were *less* repetitive in Figure 3.

Wikiplots Story

Outline: the rocky horror picture show _kw_ convention attendees also includes servants riff raff _kw_ annual transylvanian convention _kw_ old high school science teache r_kw_ frank justifies killing eddie _kw_ enraged rocky gathers frank _kw_ rainy late november evening _kw_ dr scott investigates ufos _kw_ jealous frank kills eddie _kw_ live cabaret floor show _endkw_

Article: A criminologist narrates the tale of the newly engaged couple, Brad Majors and Janet Weiss, who find themselves lost and with a flat tire on a cold and rainy late November evening, somewhere near Denton in 1974 . Seeking a telephone, the couple walk to a nearby castle where they discover a group of strange and outlandish people who are holding an Annual Transylvanian Convention . They are soon swept into the world of dr Frank-N-Furter, a self-proclaimed "sweet transvestite from Transsexual, Transylvania" . The ensemble of convention attendees also includes servants Riff Raff, his sister Magenta, and a groupie named Columbia .
In his lab, Frank claims to have discovered the "secret to life itself" . His creation, Rocky, is brought to life . The ensuing celebration is soon interrupted by Eddie (an ex-delivery boy, both Frank and Columbia's ex-lover, as well as partial brain donor to Rocky) who rides out of a deep freeze on a motorcycle . Eddie then proceeds to seduce Columbia, get the Transylvanians dancing and singing and intrigue Brad and Janet . When Rocky starts dancing and enjoying the performance, a jealous Frank kills Eddie with a pickax . Columbia screams in horror, devastated by Eddie's death . Frank justifies killing Eddie as a "mercy killing" to Rocky and they depart to the bridal suite .
Brad and Janet are shown to separate bedrooms, where each is visited and seduced by Frank, who poses as Brad (when visiting Janet) and then as Janet (when visiting Brad) . Janet, upset and emotional, wanders off to look for Brad, who she discovers, via a television monitor, is in bed with Frank . She then discovers Rocky, cowering in his birth tank, hiding from Riff Raff, who has been tormenting him . While tending to his wounds, Janet becomes intimate with Rocky, as Magenta and Columbia watch from their bedroom monitor .
After discovering that his creation is missing, Frank returns to the lab with Brad and Riff Raff, where Frank learns that an intruder has entered the building . Brad and Janet's old high school science teacher, dr Everett Scott, has come looking for his nephew, Eddie . Frank suspects that dr Scott investigates UFOs for the government . Upon learning of Brad and Janet's connection to dr Scott, Frank suspects them of working for him; Brad denies any knowledge of it, and dr Scott assures Frank that Brad is totally not involved in UFOs . Frank, dr Scott, Brad, and Riff Raff then discover Janet and Rocky together under the sheets in Rocky's birth tank, upsetting Frank and Brad . Magenta interrupts the reunion by sounding a massive gong and stating that dinner is prepared .
Rocky and the guests share an uncomfortable dinner, which they soon realize has been prepared from Eddie's mutilated remains . Janet runs screaming into Rocky's arms, provoking Frank to chase her through the halls . Janet, Brad, dr Scott, Rocky, and Columbia all meet in Frank's lab, where Frank captures them with the Medusa Transducer, transforming them into nude statues . After dressing them in cabaret costume, Frank "unfreezes" them, and they perform a live cabaret floor show, complete with an RKO tower and a swimming pool, with Frank as the leader . Riff Raff and Magenta interrupt the performance, revealing themselves and Frank to be aliens from the planet Transsexual in the galaxy of Transylvania . They stage a coup and announce a plan to return to their home planet . In the process, they kill Columbia and Frank, who has "failed his mission" . An enraged Rocky gathers Frank in his arms, climbs to the top of the tower, and plunges to his death in the pool below . Riff Raff and Magenta release Brad, Janet, and dr Scott, then depart by lifting off in the castle itself . The survivors are then left crawling in the dirt, and the narrator concludes that the human race is equivalent to insects crawling on the planet's surface, "lost in time, and lost in space .

Table 6: Example Training Story from Wikiplots Dataset.

Writing Prompts Story

Outline: found something protruding _**kw**_ geometric shapes glowing _**kw**_ sister kneeling beside _**kw**_ dead bodies everywhere _**kw**_ running around trying _**kw**_ written anything also
textbf_kw_ darkness overwhelmed _**kw**_ firelight flickering _**kw**_ could hear something _**kw**_ definitely screams _**endkw**_

Article: It was dark and Levi was pretty sure he was lying on his back . There was firelight flickering off of what was left of a ceiling . He could hear something but it was muffled . He was almost positive it was screaming . When he tried to move he felt an excruciating pain in his left side that caused him to cry out . His hand moved to it instinctively and found something protruding from the wound . It seemed to be a pipe of some kind . He attempted to pull it out but it was covered in blood making it slick . Levi 's ears began ringing and the sounds began to become clearer , it felt quite a bit like some one was driving needles into his eardrums . The sounds he was hearing were definitely screams and not just one person 's , a lot of people were screaming or yelling . There was some one close to him that was crying . He looked in the direction of the tears and seen his sister kneeling beside him , her hands covering her face . " What happened Laur ? ".

He was shocked at the sound that barely rasped out from between his lips . His sister 's hands jerked down and she stared down at Levi with a shocked look on her face . " bu ... I tho ... you were n't breathing ! " What started as a whisper ended in yell as she threw her self across her brother and began to sob anew . Levi cried out hoarsely but she did n't hear . She just continued to cling to him like a little girl that had just found her lost doll . He put one of his arms around her and scanned the room as much as he could . It looked like a warzone , like something out of one of the many shooters in his gaming collection . " What the hell ? " There were dead bodies everywhere , he recognized some of them . There were firefighters and EMT 's running around trying to find survivors in the rubble . Most of the screams were coming from survivors .

He seemed to be laying on top of the bottom of a desk , and he was pretty sure the pipe sticking out of his side was a actually one of the legs . Then he spotted it lying about two feet from his right leg , a round section of desk about the size of a softball . On it was a round symbol with geometric shapes glowing with dark red embers and a dull tendril of smoke rising up from it . It all came back to him in rush . He drew that in his notebook . It was second period and his trig teacher had this monotonous voice that could put a crack head to sleep . Laurana caught him doodling and had thrown a pencil at him to scold him silently , which made him jerk as he was connecting the last line on his drawing . Then there was the light and the heat and lastly the dark . Did he do this ? What the hell was going on here ? A voice brought him out of his little flashback . " Sir . Sir ? Sir ! "

He seemed to be laying on top of the bottom of a desk , and he was pretty sure the pipe sticking out of his side was a actually one of the legs . Then he spotted it lying about two feet from his right leg , a round section of desk about the size of a softball . On it was a round symbol with geometric shapes glowing with dark red embers and a dull tendril of smoke rising up from it . It all came back to him in rush . He drew that in his notebook . It was second period and his trig teacher had this monotonous voice that could put a crack head to sleep . Laurana caught him doodling and had thrown a pencil at him to scold him silently , which made him jerk as he was connecting the last line on his drawing . Then there was the light and the heat and lastly the dark . Did he do this ? What the hell was going on here ? A voice brought him out of his little flashback . " Sir . Sir ? Sir ! "

Table 7: Example Training Story from WritingPrompts Dataset.

NYT Article

Outline: upcoming annual economic summit meeting _**kw**_ take intermediate steps_**kw**_ says white house _**kw**_ prevent serbian aggression _**kw**_ meet boris yeltsin _**endkw**_

Article: The long-simmering tensions in Serbia's province of Kosovo turned violent in recent weeks and threaten to ignite a wider war in the Balkans. Only a concerted diplomatic effort by the United States can keep the conflict from escalating. Though he has been attentive to the problem, President Clinton must do more to take the lead with European nations to insure that Kosovo is not left adrift.

Since Slobodan Milosevic, the Serbian leader, stripped Kosovo of its autonomy in 1989, Kosovo's overwhelmingly Albanian population has engaged in mostly peaceful resistance. It brought them nothing but more repression. In recent months, an Albanian guerrilla army has emerged, targeting mainly Serb policemen. The guerrilla campaign has angered Serbs and given Mr. Milosevic an excuse to bomb villages and carry out indiscriminate attacks. He appears to be trying to push the 1.8 million Albanians out of Kosovo entirely.

A war in Kosovo, massacres of Albanians or a rush of refugees into Albania and Macedonia could bring those two neighboring countries into the conflict. It might also destabilize the fragile peace in Bosnia and flood Turkey with refugees. Even Turkey and Greece, ancient enemies, might be tempted to intervene to enhance their influence in the Balkans, especially if Macedonia is in chaos.

International responsibility for dealing with the Kosovo crisis rests primarily with the United States, Britain, France, Italy, Germany and Russia. Acting together as the Contact Group, they are trying to force Mr. Milosevic to accept internationally supervised negotiations with the Albanians. But the group has proved ineffectual because its powers are limited and some members, notably Russia, oppose strong pressure against Serbia. The group has frozen Serbia's assets abroad and this weekend imposed a ban on new foreign investment in Serbia. The sanctions, however, are impossible to enforce among countries outside the Contact Group and difficult even inside it, given Russia's views.

When President Clinton meets Boris Yeltsin later this week at the annual economic summit meeting, he should seek more Russian cooperation in pressuring Serbia. He sent a high-level delegation to Belgrade this weekend to say that Serbia will remain isolated if fighting continues. But there is little indication that Mr. Milosevic cares.

The White House has not ruled out the use of force to prevent Serbian aggression in Kosovo, but other, intermediate steps should be used before Mr. Clinton considers military action. NATO at this stage can play an important role by increasing its visibility in the region. NATO soldiers ought to be added to a peacekeeping force already based in Macedonia, and a similar group should be stationed in the north of Albania to secure the border and control weapons smuggling. But NATO should also push Mr. Milosevic to accept NATO observers in Kosovo, which he might do if he fears the guerrillas are growing too fast. If Western nations cannot muster a clear and unified message to Mr. Milosevic to restrain his army, he will unleash a new round of ethnic killing in the Balkans.

Table 8: Example Training Story from New York Times Dataset.

	baseline	% prefer PM	SEM	p-val
q1-outline	Fusion	95	4.9	0.00
q2-repetition	Fusion	55	11.1	0.67
q3-transition	Fusion	90	6.7	0.00
q4-relevance	Fusion	55	11.1	0.67
q5-beginning	Fusion	90	6.7	0.00
q6-ending	Fusion	100	0.0	0.00
q7-order	Fusion	85	8.0	0.00
q1-outline	GPT	80	8.9	0.00
q2-repetition	GPT	65	10.7	0.19
q3-transition	GPT	80	8.9	0.00
q4-relevance	GPT	85	8.0	0.00
q5-beginning	GPT	90	6.7	0.00
q6-ending	GPT	85	8.0	0.00
q7-order	GPT	85	8.0	0.00
q1-outline	GROVER	25	9.7	0.02
q2-repetition	GROVER	65	10.7	0.19
q3-transition	GROVER	65	10.7	0.19
q4-relevance	GROVER	60	11.0	0.38
q5-beginning	GROVER	65	10.7	0.19
q6-ending	GROVER	55	11.1	0.67
q7-order	GROVER	60	11.0	0.38

Table 9: Small-scale human study: H2H comparison of PLOTMACHINES (PM) with baseline output for 20 full stories. SEM is the standard error of the mean. The p-value is a t-test comparing to 50% (no preference between outputs). Although this is a small-scale study, the preference for PM is significant in many of the comparisons to Fusion and GPT2. Overall, there is a general trend towards PM being preferred in all cases except for the comparison of outline utilization with GROVER.

Model	Narrative Flow			Order
	Rep(↓)	Tran(↑)	Rel(↑)	Acc(↑)
Fusion	2.61±0.09	2.98±0.08	3.36 ±0.08	**73**±4.4
GPT	**1.39**±0.06	1.89±0.09	2.06±0.10	42±4.9
GROVER	1.78±0.08	3.00±0.11	3.29±0.11	62±4.9
PM	1.64±0.07	**3.02**±0.10	**3.39**±0.10	59±4.9

Table 10: Extended results with standard error of the mean for human evaluations of paragraph excerpts from Fusion, GPT, GROVER and PLOTMACHINES (PM) outputs. Narrative flow questions rate the repetitiveness between paragraphs, transitioning, and relevance within paragraphs.

A.3 Qualitative Examples

In this section, we include examples of model outputs on the validation set with annotations for incorporated outline points.

We show example full stories from the Wikiplots validation set comparing outputs from:

- GROVER (Figure 9) and PLOTMACHINES (Figure 10)
- GROVER (Figure 11) and PLOTMACHINES (Figure 12)
- Fusion (Fan et al., 2018) (Figure 13) and PLOTMACHINES (Figure 14)

In the examples, we highlight outline points that are mentioned in red. We also bold a few sections in the GROVER output where the model notably ends the story and starts a new one. Examples indicate that GROVER often finishes the story and then starts a new story partway through the document. This shortcoming may help explain why GROVER over-repeats outline points and why humans judge it to be more repetitive and less consistently relevant. In contrast, our models adhere more to a beginning-middle-ending structure.

We also show additional examples of introduction and conclusion paragraphs generated by PLOTMACHINES (Table 11), demonstrating the discourse the model has learned. For example, the model often starts stories by setting the scene (e.g. "In the early 1950s, a nuclear weapons testing continues ....") and often ends with a definitive closing action (e.g. "... the film ends with humperdinck and buttercup riding off into the sunset.")

Paragraph type	Paragraph
intro	in the early 1950s, a nuclear weapons testing continues at an underwater hydrogen bomb test site. scientists are concerned that it may be too dangerous to detonate without being detected by radar and radiation detectors. government sends paleontologist kyohei yamane (kim kap) to investigate. he is killed when his boat explodes while on shore patrol. as evidence describes damage consistent with sabotage of oil rigs, they conclude there must have been more than one way inside the facility. meanwhile, military research has discovered a deep underwater natural habitat alongside others where water can not be mined for life - saving purposes.
intro	the novel is set in a post - apocalyptic future where earth almost uninhabitable, with only one habitable planet for habitation and an intersolar system police force (rf) to maintain order. the story begins when " cowboy bebop ", who has been living on his homeworld of nepal since he was 12 years old, returns from space after being stranded by a comet that destroyed most of the interstellar civilization. he finds himself at home as well as friends among other characters.
intro	in 1933, joker terrorizes gotham city by murdering the mayor of new york. commissioner gordon is called to defend gotham whenever crime strikes are occurring and he has his own reasons for doing so : a corrupt police lieutenant eckhardt (james stewart) wants napier captured alive ; an elderly woman who was once part of batman 's gang tries to kill him but instead accidentally drops her gun into the water. joker also becomes obsessed with capturing the joker. meanwhile, photojournalist vicki vale begin their investigation on batman as well as other characters from the newspaper " big daddy " and " the joker ".
conclusion	humperdinck arranges for buttercup to get married to a powerful don juan carlos, who is rumored to be able to control the entire province. humperdinck secretly orders rugen to kidnap buttercup and bring her to him. rugen succeeds in kidnapping buttercup, but humperdinck kidnaps her anyway. buttercup manages to free herself and flee with humperdinck, but is captured by manuela, who accuses humperdinck of trying to keep her prisoner. humperdinck swears revenge on manuela and his henchmen, and rescues buttercup just in time. the pair head north to santa fe, where humperdinck uses his magic powers to heal buttercup 's wounds. the couple settle in a small cabin owned by mrs mccluskey, who introduces buttercup to mr smith, a blacksmith. humperdinck 's plan backfires when mr smith is attacked by apache indians, and humperdinck saves him. the film ends with humperdinck and buttercup riding off into the sunset .
conclusion	stevens and angel eyes sneak into the church hall and steal a bible. stevens opens the book and reads passages from psalms 118 to 350 bc. stevens closes the book and hands it to angel eyes. angel eyes then places stevens ' hand atop the cross and prepares to strike. stevens grabs hold of angel eyes and begs him to reconsider. stevens pleads with angel eyes to listen to reason. angel eyes makes stevens tell him why he left the confederacy. stevens tells him that he was betrayed by his mother and sister and that he needs redemption. stevens then lies and tells angel eyes that he ca n't forgive him. stevens then walks away. angel eyes watches him disappear into the night .
conclusion	in 1987, toscani meets harrison at a bar, where harrison confesses that he orchestrated the bombing of harrison 's hotel room. harrison promised justice for his friends and family, but toscani refused to believe him. harrison pleads with toscani to let him live, but toscani rejects him. toscani drives away, and harrison follows him. toscani breaks down crying, realizing that he has failed. harrison promises justice for his victims, and toscani smiles sadly .

Table 11: Example introduction and conclusion paragraph generations from PLOTMACHINES using the Wikiplots validation set.

Grover

- crossfire (film)
- keeley slowly piece together
- possible witness named ginny
- robert mitchum), concerned
- police investigator finlay
- suspected soldier relays
- steve brodie), mitchell
- one possible motive

✓: mentioned (partially) at least once
✓: mentioned multiple times

red: outline points
bold: are a few notable places where the model seems to end the story and start telling a new one (note: Grover often inserts "1:" at the beginning of stories copying a section header that frequently appears at the beginning of training data.)

1:John Keeley (Robert Mitchum), concerned for the welfare of a suspected soldier, takes him to the home of Julie King (Kay Francis), a young woman. Julie tells Keeley that she has recently broken up with an angry long-time lover. After she returns home, Keeley brings the man home .

Under questioning, the man refuses to identify Keeley. Police investigator Finlay (Macdonald Carey) does a background check and uncovers that the man is an Army Ranger. Lee Thompson (Steve Brodie), Mitchell's lawyer, takes charge of the case. With the help of Julie and his brother, Dave (John Ballou), an officer with the Rangers, the father-son pair decide to delve into the Army's most notorious past, that of James Mitchell (Clarence Muse). Keeley fears that this could become the one possible motive Behind the Lens. A visit to the man's well-guarded home and some untraceable leads them to James' buried body .

Lee, who is on leave during the investigation, refuses to help. He and Keeley slowly piece together what happened in the mysterious cabin of James' two killer during the soldier's earlier stay. Some soldiers there claimed to have observed a man abandoning a wagon. The suspected soldier relays this detail to Lee and is found murdered the next day. **1:Brian Keeley (David Nugent) is a Denver Metro cop who** has a secret obsession for witnesses. Keeley does whatever it takes to protect his friends and family from conflict and hardship in the police force. As soon as he is drafted into the army, he leaves his home city to serve a three-year tour of duty in Iraq as a turret gunner. Troubled, Keeley slowly piece together the wreckage of his life before and during the course of his war time .

As the wounds of war linger, the strain of seeing other witnesses hits the Kansas city of which Keeley is an only child, Julie (Sarah Jones) is a widow living with her parents Bob (Warren Eastwood) and Mary (Shirley Jones) and her younger sister, Cassandra (Alexandra Paul). The first time he sees her, after a stress-induced incident in which he is knocked unconscious, he accidentally shoots her in the shoulder. No one on the stand sees it, a lone witness, a possible witness named Ginny (Tara Fitzgerald), has what few witnesses witness the shooting because she is the only one who had the car and gun key that made the shot. A new man at the stand is also a possible witness because the man, Stan Meredith (Steve Brodie), Mitchell (Robert Mitchum]), concerned to see that the accused soldier is not convicted before being removed from the stand. Mitchell has phone conversations with Meredith about Meredith's possible involvement in the incident. As Keeley overhears these conversations, he asks Julie to hide her purse from the police. Once Julie realizes that Keeley is searching her purse for his wallet, she is furious and gets into a fistfight with Keeley, after the two meet again in the brief moments after Keeley surprises her with the money from the purse. Keeley then goes on the stand and answers the last few question from police investigator Finlay (Harry Andrews) .

When he asks Julie to drop the fight because of his feelings for her, Julie tries to tell him that she does not feel safe with him any longer. After trying to get Karen, his pregnant ex-wife, to tell her that he is waiting for her to give birth, Keeley is attacked in the street and caught in a traffic accident with the police arriving in mere seconds. As the soldier is about to take Keeley into custody, Kelly (Trish McEvoy), a navy sailor in a bunk bed with Keeley, shoots him in the leg. Kelly leaves him lying on the ground in pain, then turns around and runs to catch her man before being run over by a truck. Keeley is taken to the hospital where Mary and Cassandra nurses him through his injury .

A newly hospitalized and recovering Keeley (Tim Kelleher) tells Julie that Kelly was responsible for his wounds and that he wants her to understand what Kelly meant. As he is recovering, Julie does not remember what she said during their encounter but realizes that she wanted Keeley to know why Kelly killed him. Katie then rushes to Julie and begs her to remember what she said. And they do. Kelly admits that, on the day that Keeley died, he had been obsessed with witnesses all his life. He even attempted to shoot Julie as a child after beating her mother. Keeley responds that he cannot see into Kelly's mind, that he knows what Kelly thinks. The film ends with Kelly relaying the details of his final minutes to Keeley. Later in the film, Kelly and Katie reconcile their differences and agree that the events that have transpired will help them look at their past more positively .

After Keeley recovers from his injury, the movie shifts focus to the investigation on the murder of Kelly Meredith. In the interrogation room at the hospital, the suspected soldier relays the name of the woman his partner killed. Julie is pulled away by Finlay but when she returns, she tells him that the names are now Murray Fenner and Wallace Johnson. In their conversation the suspect, who we later learn is Penny Rosen, proudly explains that she knows all about the death and is above suspicion. **1:The film begins with an absent mother.** Shirley (June Allyson), reading from a letter from a suspected soldier relays the report of what may have happened to the soldier's family during the Siege of Fort Benning, Georgia. The soldier was slain by sniper while home on leave. The sniper is revealed to be the son of an old friend, Tim Kegee (Eric Johnson) .

While hearing about the suspected soldier's death, police investigator Finlay (Jack Butler) and Shirley realize that the shooting must have been done by Kegee's childhood friend, Steve Brodie .

As the police slowly piece together the case, they discover that Kegee's father was once a convicted murderer. In the meantime, Steve Brodie has given the information on the shooting to detectives Finlay and Detective Shelly (Linda Darnell) .

A possible witness named Ginny (Ann Wilson) who was the soldier's wife is found murdered during the night. Finlay and Shelley (Sarah Miceli) are trying to piece together who the soldier's brother was and why the brother killed the soldier's family. One possible motive Engine Dave (Robert Mitchum), concerned about the sincerity of Kegee's father in denying his involvement with the shooting, asks to see Kegee's biological father. Kegee's father won't tell him that he is the brother of the soldier he killed .

Eventually, all evidence points to Kegee being the killer. Finlay finds out that Kegee was planning to murder the soldier's wife and daughter in order to get the money the soldier sent to Kegee's father to fulfill a promise he made him, with the promise to the latter to avoid retaliating. Kegee is ultimately arrested and convicted for the murder of his own family. **1:A suspected soldier relays** his instructions to a sniper on a smallS. Army base. Though the soldier is dead, authorities find his rifle at the scene. Evidence in the bunker points to the sniper being Don Keeley (played by Patrick Warburton), who, along with his partner, Max Cooper (played by Steve Brodie), Mitchell, one of the soldiers, and a possible witness named Ginny (played by Phyllis Threlfall) being able to identify Don Keeley. Keeley's sister, Vicky, supports her brother's actions and, when questioned, informs the police investigator Finlay (played by William McMichael) that Don Keeley has been having suicidal tendencies .

In the meantime, the squad's commanding officer, Sergeant Walker (played by Robert Mitchum), concerned that Don Keeley's "suicide" may have been real, brings in the sergeant's wife, Kathy (played by Elizabeth PeÃ±a), a detective. The sergeant and Kathy slowly piece together the case and figure out what Don Keeley had been planning. During a lull in the sniper's activities, the sergeant and his wife discover the sniper is missing, and have the sniper's comrades and Joan Bennett (played by Jannine Scott) seek to hide him out. When the sergeant realizes that a defenseless Joan has been trampled to death by his comrades and that the killing was done by someone he knew personally, he knocks her out .

In a final confrontation, Joan's friends and Joan's bodyguard, Larry Manett (played by Robert Hendley), attack Joan's husband, who is wounded, severely injured, while telling a man, Max Cooper, that he will kill Joan because of her innocent loss and revenge. While Don and his comrades and Joan's bodyguard proceed to her house, the sergeant runs to find Kathy, severely injured. He shoots a gun at Joan's house and hears shots, turns around and finds his wife dead. A final shot shows Don and his comrades running and turning to one possible motive Afterwards, the sergeant reveals that "allison" is his nickname and that if he did kill Joan, it would not be the first time he had killed a woman. He further says that Don Keeley is not a murderer, as the sniper who shot Don Keeley, is not in the house. Don ends up walking away, evading the sheriff's deputies, evading police, and disappearing into the woods.

Figure 9: Example document generated using Grover

PlotMachines

- crossfire (film)
- keeley slowly piece together
- ✓✓ possible witness named ginny
- ✓✓ robert mitchum), concerned
- ✓ police investigator finlay
- ✓ suspected soldier relays
- ✓ steve brodie), mitchell
- ✓✓ one possible motive

red: outline points
✓ : mentioned (partially) at least once
✓ : mentioned multiple times

police investigator finlay (robert mitchum) , concerned about the possible witness named ginny (susan ladd) who was found dead in a park with her throat slashed and an apparent suicide . he is also investigating the disappearance of two other men from his precinct : sergeant david mitchell (steve brodie) , mitchell 's best friend ; and detective james " mac " o'rourke (nigel maclachlan) . when scotland yard receives a tip that one possible motive for murder would be a suspected soldier relays to them via radio .

duncan begins having nightmares about ginny 's death , and eventually realizes that someone else is behind the murders . the killer then kills another person , leaving only one possible motive for the crime : the body of a teenage girl . the murderer uses a gun used to kill the victim to hide the fact that he or she committed the killing . however , the real culprit is revealed to be myra , who had earlier killed herself because she believed that she did n't deserve what happened to her . she confesses to all these crimes , including the rape and subsequent shooting of a child .

nick 's suspicions lead him to discover that the accused were members of the military who participated in the war effort . he tracks down the last member of the group , lt joseph wilson (brian heffley) , who admits to knowing the truth about the bombing . he reveals that he knew nothing about the plot , except that it involved a possible witness named ginny . nick confronts her , demanding answers . she refuses to answer , claiming that she does not remember anything about the night of the bomb blast .

van alden is forced to reveal that he is actually part of a double squad led by sgt . donovan (robert mitchum) , worried that the soldiers might be looking for evidence that could implicate him in the massacre . van alden is able to identify a possible witness named ginny (laura moore) , who was shot in the head shortly prior to the ambush . van alden 's team manages to track down the shooter , who turns out to be the same man who murdered marlene . van alden and laird colvin (dan holloway) arrest the sniper , but he escapes . van alden is ordered to bring ginny to court , where she pleads guilty to three counts of attempted manslaughter . van alden reluctantly accepts the plea bargain , though he still feels betrayed by rusty .

the rest of the story deals with the search for penny , and the resulting confrontation between rafe cordova (robert mitchum) , concerned about the possibility of penny 's safety , and the eventual conclusion that rafe cordova is innocent . the ending shows penny and rafe traveling through mexico , stopping at various places in arizona , texas , california , louisiana , nevada , and arkansas , searching for clues to the identity of the actual robber . the final scene shows penny and rafe riding off into the sunset , presumably heading for san francisco .

Figure 10: Example document generated using PLOTMACHINES

Grover

- ✓✓ cops (film)
- ✓✓ police officers throughout
- ✓✓ constructed short films
- ✓✓ national film registry
- ✓✓ united states library
- ✓✓ pork pie hat propped

✓: mentioned (partially) at least once
✓: mentioned multiple times

red: outline points
bold: are a few notable places where the model seems to end the story and start telling a new one (note: Grover often inserts "1:" at the beginning of stories copying a section header that frequently appears at the beginning of training data.)

1:The film's plot traces the fall of a Mafia don, Joe Crowther, and his relationships with other mobsters and police officers throughout the period from his conception in 1977 to his arrest in 1986 .

The official synopsis of the film is as follows: "An actor with a Pork Pie hat propped ullivan on a park bench in Manhattan in November 1982. A scowl on his face. – and a laugh". In the opening scene, Crowtherâs wife and daughter are shown to have been running a health food store in the Bronx. Crowther has a wife and daughter in Virginia as well as a son and daughter in Brooklyn. He frequently visits their home while on probation .

He is inducted into the NYPD, where he settles into his new role and soon finds himself the target of a political scandal. The crew of detectives working with Crowther, in addition to an FBI agent who works with him and his family, make use of all methods they can imagine to ruin his reputations. The film is a series of scenes about him interacting with reporters and his family. The quality of his dialogue and conviviality are characteristic of this actor, and in many of the self-competence that he displays in most of his acting jobs, his voice has been dubbed for them. This is to say that he is writing dialog to archive footage from built-in short films, and he is both speaking and laughing. The character in the film is the system .

The federal agents that stand in Crowtherâs way are his former assistants, culminating in a dinner where Crowther mocks his chief rival, that is the current don. Crowther is in for some roughhousing as the folksy talk turns into a quick game of Narcotics with the mobsters, leading to a midnight meeting in the bathroom with all of the buildingâs law enforcement. Crowther goes to sleep that night. **1:Bob Garrett, Gil Wood, Lisa Lee, and Angela Calderon, all** of whom work at the United States Library's National Film Registry, talk about how they became the "cops" of the documentary film "Police: A Formic Tale" .

They first imagined the idea of a documentary after one of their co-workers, Gilbert, commits suicide. Because Gilbert was obsessed with a horror film, which he attended regularly, he was ostracized by his co-workers. They then found a homeless man with a pork pie hat propped upside down on the sidewalk. They decided that using Bob's trailer, and Gilbert's drinking abilities as a source of budget for their project, they could use a homeless man for a second "cop" to patrol the streets of Los Angeles. Although they did not have experience, they learned about responding to incidents in a three-day period. In their film, they filmed all of the police officers throughout the following day and night, and then they stored a replica of an explosive and a coffin for their "cop" to use in a three-day period. Then all of the police began their crime-fighting duties, leading to comic incidents, a wave of violence, and tense moments. **1:The film,** entitled Police Officers Throughout the Homeland, is dedicated to the police officers who have worked and worked hard to keep order and security in communities throughout the United States .

It is a collection of nine short films which are constructed short films of the type featured in the National Film Registry of the United States Library and Archives. Each episode contains a fictional story that is the basis for the film. The films, each one focusing on a particular type of police unit, begin with a description of the unit's or the members of their unit's duties and duties in and outside of their duty station. They are then followed by the police officer or members of the patrol team at that particular unit's location. As each episode progresses, it is revealed that the units members are being under false pretenses. Inevitably, the unit is disrupted by trouble makers and other groups who have emerged from the abuses of the United States' government and by the police officers themselves. The film ends with a news segment stating that "the unit charged with protecting America's cities is none other than the Police Officers Throughout the Homeland". At that point, we see a Pork Pie Hat propped Outside Atlanta's federal courthouse. The film ends with the cartoon ending with a caption that reads: "Watch the coming of age film you'll never see. **1:The film is a series of simple**, self-constructed short films, in which a murder occurs, and how it is done is shown by the player. They include clips of police officers throughout the film, men dressed in blue suits and blue suits who claim to be cops, and a people's assembly on a bridge. The narrator describes the short as "satiric" and frequently mentions that "those with guns are always gonna shoot" .

The film can be viewed on YouTube or in the National Film Registry of the Library of America. It was produced byproductively rendering the word "police" into the words and characters (a police badge for "officers") .

According to its name tag on the cover photo, the short film is entitled "Good Cop, Bad Cop, TV Show" and "The Banks Behind The System". (1970).

Figure 11: Example document generated using GROVER

PlotMachines

- ✓✓ cops (film)
- ✓ police officers throughout
- ✓ constructed short films
- ✓ national film registry
- ✓✓ united states library
- pork pie hat propped

red: outline points
✓ : mentioned (partially) at least once
✓ : mentioned multiple times

the film is set in a united states library , located at an old - fashioned desk and surrounded by bookshelves . it consists of three stories with police officers throughout each story . there are two versions : one for each floor ; the other for every room . the narrator (played by john wayne) explains that he has constructed short films to help people who need them . these include " cops " , " policemen " , " robbers " , " criminals " , " murderers " , " killers " , " terrorists " , " gangsters " , " assassins " , " revolutionaries " , " anarchists " , " traitors " , " villains " , " murderers " , and " rapists " .

they have erected a large statue of napoleon , but when questioned why they did n't capture him yet , they simply say that their boss wanted to make sure that his men would never find out where he hid the money . after hearing the news , they decide to go undercover and infiltrate the bank owned by the president of the state department . while infiltrating the vault , they discover that the robbery will take place over a long period of time . during the heist , the group discovers that the leader of the gang is actually a member of the national film registry , known only as the chief of security . the head of the operation is named mr hoover , and the rest of the team are called " cops " .

the plan works perfectly until the thieves break into the museum using a concealed tunnel . they steal all the books and papers , leaving the prisoners free to escape . the prison guards chase after them , but the boys manage to hide themselves under desks and chairs before escaping . the boys return home to tell their parents about the incident . their mother is shocked and angry , believing that her sons ' actions led to the death penalty . she scolds them for running away without telling anyone . they explain that they just ran away because they could no longer stand living in fear .

finally , the thief reveals herself as mrs robinson , the librarian of the united states library . she informs the heroes that she knows everything about them and even helped them escape . she wants to show them something . she shows them a picture of the famous hero , william shakespeare , who died fighting the nazis . the four friends wonder what happened to him . the policeman returns and takes the photo . the others realize that the policeman is none other than william shakespeare himself . the policeman claims that he killed shakespeare because he knew too much information about the nazi regime . he leaves .

the film features a detailed description of the structure of the library , including its construction and layout . the illustrations are based upon actual events . the buildings featured in the pictures depicted are modeled after those found in real life such as stonehenge , atlantis , mount rushmore , the great pyramid , etc . the photographs depict the entire town of granville , california , and the surrounding countryside . the map used in the documentary is described as having been taken from 1899 to 1947 . the location of granville is shown in the film .

Figure 12: Example document generated using PLOTMACHINES

Fusion

- the hound of the baskervilles
- james mortimer asks sherlock holmes
- better even inside baskerville hall
- ✓ strange sightings trouble watson
- friend sir charles baskerville
- sir hugo baskerville abducted
- ✓✓ baskerville family since
- ✓ bearded man roaming free
- watson grows suspicious
- many baskerville heirs

red: outline points
✓ : mentioned (partially) at least once
✓ : mentioned multiple times
bold: examples of repetitive/degenerate output

Sherlock Holmes , **Watson , Watson , Watson , Watson , Watson , Watson , Watson , Watson and Watson , Watson , Holmes , Watson , Watson , Watson , Watson , Watson , Watson , Watson , Watson , Watson and Watson** . Watson is a <unk> who has been working on a mysterious , <unk> Moriarty , who has been on the hunt for the Baskerville family since his father died . Watson has been investigating the murder of a masked man roaming free of the Baskerville family from Scotland . Watson is a <unk> who has been searching for the secret of a strange sightings of the Baskerville family . Holmes has been trying to retrieve the missing key from Watson ' s father , **who has been killed by the man he killed** .

Holmes ' son , Watson , is now on the trail of the dead man 's body , a ghost who has disappeared . Watson comes to Watson ' s office with the help of Watson , who has disappeared . **Watson finds Watson and Watson** , who have come there . Watson , who has a crush on Watson

Figure 13: Example document generated using the Fusion model (Fan et al., 2018)

PlotMachines

- the hound of the baskervilles
- ✓✓ james mortimer asks sherlock holmes
- ✓ better even inside baskerville hall
- strange sightings trouble watson
- ✓ friend sir charles baskerville
- ✓✓ sir hugo baskerville abducted
- ✓ baskerville family since
- ✓ bearded man roaming free
- ✓ watson grows suspicious
- ✓✓ many baskerville heirs

red: outline points
✓ : mentioned (partially) at least once
✓ : mentioned multiple times

holmes and watson are summoned to the baskerville family mansion in england by sir hugo baskervilles , who has been missing for several days . he is found dead with his throat slashed . it appears that a wild animal had killed him . inspector mortimer asks sherlock holmes if there was any connection between this murder and baskerville 's disappearance . holmes replies that there were many baskerville heirs but none of them could have done so without being seen or heard from again . they also find out that baskerville 's friend sir charles baskerville abducted his daughter susan and her husband george .

holmes suggests that the case might be linked somehow to the mysterious death of a young girl , charlotte barrymore , whose body was discovered near the moor during the night before her abduction . however , the police suspect that the murderer must have been one of many baskerville heirs . holmes agrees to investigate further , and soon learns that the victim was murdered by a large group of men dressed like monks . these include the hunchbacked brother - abbot osmond , the old priest , and the sinister , elderly bishop .

holmes suspects that the murders are connected to the notorious black knight , the earl of winchester , who reportedly disappeared mysteriously over 1870 . the investigation leads to the infamous count de mornay , known only as the baron von lestrasse , who vanished just prior to the start of world war ii . the baroness of ravensburg , countess catherine treherne , has recently returned to england from france , accompanied by her nephew , prince edward iii , heir apparent to the throne . she plans to marry off her son to the marquis de sade , viscount anthony grafton , and take control of ravensburg .

holmes investigates the circumstances surrounding the kidnapping , finding evidence that links the victims to various other families , such as those belonging to prominent businessmen , clergyroppers , and bankers . he comes across a bearded man wandering free in the woods , who identifies himself as nicholas mather , a relative of the deceased patriarch of baskerville hall . mather claimed to have met a stranger in paris , claiming to have witnessed a violent struggle between two assailants , both of whom were wearing masks similar to those worn by mather . mather claimed that he saw no witnesses , nor did anyone else see anything .

mortimer seeks advice from sir hugo baskerville , a retired detective , who advises him to seek out the elusive baskerville . baskerville has long lived in seclusion in the mountains , haunted by nightmares caused by wolves . mortimer hires a private investigator , max walker , to follow baskerville , hoping to learn whether he is still alive . walker warns mortimer that baskerville is indeed alive , but is dying . mortimer grows suspicious of baskerville 's behavior , especially when he hears noises outside the walls . mortimer eventually manages to track baskerville down , and finally confronts him . baskerville admits that he is insane , and insists that mortimer kills him . mortimer shoots baskerville , but misses and wounds the wounded baskerville . mortimer flees , pursued by walker , who chases him into the forest . mortimer recovers his revolver and pursues walker , who eludes him once again . mortimer reaches the edge of a cliff overlooking the sea , where he encounters a shadowy figure standing above him . mortimer fires twice , hitting the figure in the chest . mortimer falls to his knees , clutching his wound . walker stands over him , holding his hand , implying that mortimer has mortally injured him . mortimer dies , apparently relieved .

Figure 14: Example document generated using PLOTMACHINES