NAREOR: The Narrative Reordering Problem

Varun Gangal * 1 Steven Y. Feng * 1 Malihe Alikhani 2 Teruko Mitamura 1 Eduard Hovy 1

1 Language Technologies Institute, Carnegie Mellon University
2 School of Computing and Information, University of Pittsburgh
{vgangal,syfeng,teruko,hovy}@cs.cmu.edu, malihe@pitt.edu

Abstract

Many implicit inferences exist in text depending on how it is structured that can critically impact the text’s interpretation and meaning. One such structural aspect present in text with chronology is the order of its presentation. For narratives or stories, this is known as the narrative order. Reordering a narrative can impact the temporal, causal, event-based, and other inferences readers draw from it, which in turn can have strong effects both on its interpretation and interestingness. In this paper, we propose and investigate the task of Narrative Reordering (NAREOR) which involves rewriting a given story in a different narrative order while preserving its plot. We present a dataset, NAREORC, with human rewritings of stories within ROCStories in non-linear orders, and conduct a detailed analysis of it. Further, we propose novel task-specific training methods with suitable evaluation metrics. We perform experiments on NAREORC using state-of-the-art models such as BART and T5 and conduct extensive automatic and human evaluations. We demonstrate that although our models can perform decently, NAREOR is a challenging task with potential for further exploration. We also investigate two applications of NAREOR: generation of more interesting variations of stories and serving as adversarial sets for temporal/event-related tasks, besides discussing other prospective ones, such as for pedagogical setups related to language skills like essay writing and applications to medicine involving clinical narratives.

1 Introduction

From the onset of language, storytelling has been crucial to the transmission of knowledge (Ramanujan 1991). It has been well-established that readers remember only an abstract representation of stories (Schank 1972). Before the printing press, classes engaged with oral teaching of scriptures, such as rabbis, underwent extensive training to reproduce them with no distortion (Bos 1995). Formally analyzing story structure commenced with the ancients, through works like Aristotle’s Poetics (Halliwell et al. 1998). These studies led to the concept of a narrative, distinct from story events.

For a story, there are two orders: the chronological order of events as they happened and their order as presented in text. These have been analyzed under different names (Propp 2010). We refer to them as story order and narrative order, or story and narrative, respectively. Genette (1983) enlists typical orders observed in writing. A linear order narrates events in same sequence as story order. The in medias res order starts with events in the middle, goes back to the start, then proceeds to the end. Changing from near-linear to more “interesting” orders is prevalent in cinema, e.g. The Imitation Game starts with Turing’s post-WWII 1951 interrogation. Memento and Naked Lunch are known for their esoteric narrative orders - loosely described as retrograde (reverse of linear) and syllepsis (lacking chronological logic), respectively.

Morgan (2017) explains how narratives surpass “mere chronicle”. Narrative orders of presenting materials in scientific explanations directly affects how researchers interpret and understand them since the order implies not only temporal but other inferences about causality, processes of change, etc. Narrative order can thus influence model explainability, especially for explanation generation (Rajani et al. 2019), a recent area-of-interest (Wiegreffe and Marasovic 2021).

In this work, we do not delve into the complex and somewhat subjective question of which narrative order is most suitable or “interesting”. We focus on how a given story in linear narrative order can be rendered in a specified, non-linear, target order while preserving plot. We call this Narrative Reordering, or NAREOR. To the best of our knowledge, we...
are the first to propose and investigate this task.

Our work is not entirely adrift from past research in this vein. Montfort (2007) tries generating fiction narratives from basic existent-event info with a special focus on narrative order, using a rule and planning approach. Unlike our work, their rule-based system does not involve learning. Moreover, being generation in a given narrative order from unstructured story elements rather than reordering an existing story, their setting does not require solving challenges such as disentangling events from stories which are inherent in NAREOR.

Formally, NAREOR involves reordering a story \( S \) with sentences \( s_1, s_2, \ldots, s_n \) to a reordered story \( S' \) with sentences \( s'_1, s'_2, \ldots, s'_n \) according to a given target narrative order \( \pi \). \( \pi \) is a permutation \( \pi : i' \to f(i') \) \( 1 \leq i' \leq n \) mapping from target sentence indices \( i' \) to original sentence indices \( i \), where \( f \) is a one to one and onto function from \( \{1, 2, \ldots n\} \) to itself. In practice, we write \( \pi \) as the sequence \( \{i = f(i') \}_{i'=1}^{i'=n} \) (\( f \) and \( i' \) become implied).

NAREOR's challenges are evident from the example in Figure 1. Simply reordering sentences is far from sufficient, as rewritten text must be adjusted to handle coreference, tense, and other discourse dependencies. For example, narrative order affects tense since it can change the first person pronoun: examples of the sentence (e.g. "I like to earn enough money to take some time to travel") are rewritten to clarify the now disrupted flow. We find changes in occurrence distribution across verb-argument and other discourse dependencies. For example, narrative order dependent elements like ellipses (e.g. using NLTK's Gerund fraction (e.g. "I like to earn enough money to take some time to travel").

2 Dataset: NAREORC

2.1 Dataset Construction

Source Corpus: ROCStories has \( \approx 98.5K \) five-sentence English stories. For the dev and test split, each example contains a four-sentence story prefix with a one-sentence coherent and incoherent ending. We treat the coherent endings as the fifth sentences for NAREORC’s dev and test stories.

Assigning Target Narrative Orders: The target narrative order \( \pi \) is not part of the ROCStories input. We devise a randomized procedure to assign a reasonable \( \pi \) for each example. We sample 3 permutations from the set of non-identity \( n!-1 \) permutations. We find Kendall \( \tau \) correlations (Kendall 1938) between identity permutation \( I_n \), \( \{1, 2, 3, 4\} \), and each of the three permutations, retaining the lowest as \( \pi \). We prefer this to sampling at random because we want our examples to be sufficiently non-trivial w.r.t. the task.

Supervised & Unsupervised Splits: We set aside 600, 200, 200 stories from train, dev, and test splits of ROCStories. These act as NAREORC’s trainSup, devSup, and testSup splits, for which we collect human references. Remaining stories in each ROCStories split are retained as trainUnsup, devUnsup, and testUnsup of size 95161, 1671, 1671.

Human Annotation: For trainSup and devSup, we annotate one reference per example. For testSup, we collect two each to help reference-based metrics. We conduct our study on AMT. To understand task difficulty, we ask a “Hardness” question with options VeryEasy, Easy, Moderate, Hard, Very-Hard. On average, annotators found \( \approx 70\% \) of rewritings to be Moderate or Hard, demonstrating that NAREOR is quite difficult even for humans. More details in Appendix B.

2.2 Dataset Analysis

Overall Statistics

We find human-rewritten stories \( S' \) are \( \approx 1.2x \) as long as input stories \( S \) on avg in words and characters. We expect this given the narrative reordered story favors resolution of sentence-order dependent elements like ellipses (\( s_3 \) and \( s'_3 \) in Figure 1) and pronouns (\( s_3 \) and \( s'_3 \) in Figure 1) to explicit forms. It also requires insertion of time expressions (e.g Before that, 3rd row, Table 1) to clarify the now disrupted flow.

Unique n-gram ratio \( UR_n(S) \) is the fraction of unique n-grams of length \( n \) in \( S \). We observe all three mean URs \( n \) \( (n=1,2,3) \) to decrease from input to reference story. \( UR_1: \approx 6.92 \to \approx 6.69 \), \( UR_2: \approx 9.40 \to \approx 9.31 \), \( UR_3: \approx 9.89 \to \approx 9.84 \). Increased n-gram repetition could have reasons similar to length increase, causing cross-sentence repetition. Figure 1 demonstrates this: \( S \) only has one instance of money. Conversion of inherit any of it (\( s_3 \)) to inherit any of the money (\( s'_3 \)) and enough to take time (\( s_4 \)) to enough money to take some time (\( s'_4 \)) among other changes, results in four in \( S' \).

How Verb Forms Change

We note changes in occurrence distribution across verb-related \( pos \) tags from \( S \) to \( S' \) using NLTK’s \( pos \) tagger. Gerund fraction (\( pos=VBG \)) (e.g. I like playing) increases \( 7.7\% \to 9.5\% \). Past participle fraction (\( pos=VBN \)) (e.g. He had broken it) \( \approx \) doubles, \( 6.5\% \to 12.4\% \). Past tense fraction (\( pos=VBD \)) (e.g. He broke it) decreases \( 60.9\% \to 54.6\% \).

In our case, \( n = 5 \) as we experiment with ROCStories.
Other verb-related pos fractions remain fairly constant. Increase in past participle can be explained by frequent conversion to past perfect tense during reordering (e.g. parents passed away → parents had passed away in Figure 1).

How Narrative Reordering Alters Sentences
We look at corresponding sentence pairs \{s_i, s'_i\} in each story, specifically 4 linguistic change types - ellipsis, tense, time expressions (timexes), coreference. We tried detecting these using off-the-shelf tools, and did not find any for ellipsis. Timex detectors like SUTime (Chang and Manning 2012) only mark strict timexes (e.g. last Sunday) but not others (e.g. before midterms). We hence hand-annotate these four for each \{s_i, s'_i\} per testSup example. These are further described in Table 1. We find over half (51.5%) the examples show ≥3 of 4 change types at once, and 89.5% show ≥2. This shows that NAREOR requires performing different changes in tandem.

3 Methodology

3.1 Training Methods
We introduce two task-specific training methods.

NAR-denoise (NAR-d)
This is partially inspired by how humans rewrite; a common approach is to first reorder sentences naively (simply swap positions), then make other changes. NAR-d attempts to mimic this, learning to convert from naive reorderings to high-quality text. It involves two stages of model training.

1. Denoise-1S: Stage 1 is unsupervised training through story-level denoising. We use trainUnsup without human-written reorderings, and simulate them using the original human-written ROCStories (the outputs during training). Deletion and swapping of tokens are used to create inputs from these stories that simulate naive reorderings. This noise-hinging aims to emulate the reverse of the content editing that occurs during NAREOR. Specifically, we randomly delete 12.5% of tokens and swap another 12.5%. We found human-written stories were, on average, in combination of token length (longer) and swappings, ~25% different from the originals. We split this between deletion and swapping to approximate naively-reordered stories. Story sentences \(s\) are first reordered as per \(\pi_{iv}\) to produce \(S'_{naitve}\), then each is edited to fit the new narrative. We swap tokens as humans often swap words like coreferent mentions based on how the narrative order changes. Hence, this stage learns to denoise text by converting nosed versions to human-written text.

2. Denoise-2S: The second stage is supervised training atop the model above. The inputs are the 600 original stories in trainSup, with sentences naively reordered as per target narrative order \(\pi_{iv}\) to \(S'_{naitve}\), and the outputs are the human rewritings of these. The model learns to further translate from naively-reordered text to fluent human-written text.

NAR-reorder (NAR-r)
Unlike NAR-d, NAR-r models themselves handle reordering given the target order rather than naive reordering beforehand.

- Input Encoding Scheme: We describe how the task input \{S, \pi_{iv}\} is encoded as a token sequence for both Stage-1 and 2 training. To enable the model to distinguish different sentences, we prefix each \(s \in S\) with a tag from \(<a>\) to \(<e>\). We specify \(\pi_{iv}\) as a sequence of these, separated from \(S\) by \(<sep>\). NAREOR involves rearranging mention types among coreference chains (see §2.2), so we use NeuralCoref (HuggingFace 2020) to detect these chains. For each, we assign a unique uppercase tag \(<X>\) to replace its mentions. At the end of the input, we list each tag and the head mention of its coreference chain in order. We then append \(<st>\) to mark the end of the input. An illustration of the scheme follows: \(<a>\) Since I had front seat tickets, I was able to directly see \(<XI>\). \(<b>\) \(<XI>\) tried to reach out with \(<X2>\). \(<c>\) I grabbed \(<X2>\) and \(<XI>\) pulled me on stage. \(<d>\) \(<XI>\) began to sing. \(<e>\) the concert had started. \(<sep>\) \(<e>\) \(<a>\) \(<b>\) \(<c>\) \(<XI>\) The music artist \(<X2>\) her hand \(<st>\).

- Reorder-1S: We use examples from trainUnsup for stage 1. It is problematic to train for the forward direction of our task \(S, \pi_{iv} \rightarrow S'\) since \(S'\) is not known. Approximating \(S'\) using \(S'_{naive}\) would hurt output fluency. We instead train in the inverse direction \(S_{naive}, \pi_{iv}^{-1} \rightarrow S\), where \(\pi_{iv}^{-1}\) is the inverse permutation of \(\pi_{iv}\). To reduce train-test mismatch, we use the inverse formulation half the time, and an autoencoding one, i.e. \(S, I_n \rightarrow S\) the other half.

- Reorder-2S: trainSup examples are used to further finetune on reorder-1S. We train in the task direction \(S, \pi_{iv} \rightarrow S'\).

3.2 Chosen Models
We choose several pretrained generation models: GPT-2, BART, and T5. We finetune all using both our training methods to produce denoise-1S (d-1S), denoise-2S (d-2S), reorder-1S (r-1S), and reorder-2S (r-2S) versions. GPT-2 (Radford et al. 2019) is a Transformer-based language model trained on WebText. BART (Lewis et al. 2020) and T5 (Raffel et al. 2020) are Transformer seq2seq models. BART is trained as a denoising autoencoder to reconstruct original from noised text. T5 is designed to be effective for transfer learning. We use HuggingFace’s implementations of their base versions.4

3.3 Automatic Evaluation Metrics
Reference-Based Metrics assess the similarity between generated text and human-written references. We use BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and BERTScore (Zhang et al. 2019). We compare generated text with the two references per testSup example.5

Target Order Fidelity (TOF) is defined as how closely the reordered text matches the given target narrative order. E.g. given \(S = \{s_1, s_2, s_3\}, \pi_{iv} = \{3, 2, 1\}\), and \(S' = \{s'_1, s'_2, s'_3\}\), we wish to see if \(s_1\) has correctly been translated to \(s'_1\). We introduce TOF-METEOR and TOF-BERTScore. These assess the average METEOR and BERTScore values for each aligned pair \(\{s_i, s'_i\}\) \(\forall i\) (where \(i'\) refers to the target index for \(s_i\)). Higher values correspond to more content preservation, where each output sentence is more likely in the correct position. Some drop is expected in modulating for \(\pi_{iv}\), but the overall content should be faithful. These metrics serve more as validation, where reasonable values (e.g. > 50)6 are

4See §4 for further training/finetuning details.
5Correlates well with human evaluation as shown in §5.
6Assuming the values are multiplied by 100.
Ellipsis
(Sent: 5.7%)
(Stor: 27.5%)

S: 1. All of the Ross family has red hair, except Henry. 2. Henry has blonde hair that is very curly. 3. Henry’s father often teases Henry’s mother about the mailman. 4. The mailman has blonde, curly hair, but he is very ugly. 5. His dad’s teasing makes Henry feel bad. ; πv: {1, 5, 2, 4, 3}

S’: 1. All of the Ross family has red hair, except Henry. 2. His dad’s teasing about the mailman makes Henry feel very bad. 3. This is because the mailman has blonde, curly hair, but he is very ugly. 4. Henry also has blonde hair that is very curly. 5. Henry’s father often teases Henry’s mother about the mailman.

Tense
(Sent: 19.1%)
(Stor: 64.0%)

S: 1. Sam bought a new SUV. 2. It was all wheel drive. 3. He figured he would take it off road. 4. He hit a few hard bumps and broke his suspension. 5. Sheepishly, he brought it to the dealership for repair. ; πv: {2, 3, 5, 1, 4}

S’: 1. Sam’s SUV was an all wheel drive. 2. He thought he could take it for a spin off road. 3. Embarrassed by the outcome of his drive, Sam took the car to the dealership for repair. 4. He had just bought the SUV. 5. The car had hit a few hard bumps and the suspension broke when Sam took it off road.

Timexes
(Sent: 34.0%)
(Stor: 85.5%)

S: 1. There was once a kitten that did not have a home. 2. The poor kitten walked around cold and hungry. 3. One day, a nice lady let the kitten into her home. 4. The woman gave the kitten food and a bed. 5. The kitten was happy to be adopted.

S’: 1. A woman gave a home to a cat. 2. Before that it was cold and hungry. 3. It made the cat happy to have a home. 4. The little cat originally was homeless. 5. But in the end, it met the nice woman and she let it in.

Coreference
(Sent: 20.7%)
(Stor: 71.5%)

S: 1. Jimmy wandered around the city looking for a place for a soda. 2. Before he knew it, he was in an unfamiliar area. 3. He was scared of strangers and didn’t want to ask anyone. 4. Soon a policeman came by and asked if he was lost. 5. He told him that he was lost. ; πv: {5, 4, 2, 1, 3}

S’: 1. Jimmy told a police officer that he was lost. 2. He was lucky the police showed up in the first place. 3. He had no idea where he was. 4. He had wandered off when trying to find somewhere to buy a soda. 5. It was pretty terrifying being all alone in a mysterious area with strangers.

Table 1: Sentence pairs in testSup stories are annotated for 4 linguistic change types common in NAREORC. Sent denotes % of sentence pairs showing that change type. Stor denotes story pairs (S, S’) where ≥ one sentence pair shows that change type.

4 Experiments

Model Finetuning and Generation

For finetuning our models, we try different combinations of learning rates (LR) for both stages. We look at either the loss (for BART and T5) or perplexity (for GPT-2) on the respective validation splits (devUnsup for 1st stage and devSup for 2nd), and choose the epoch with the lowest.

We evaluate each model on testSup, where we can directly compare results to NAREORC’s human rewritings. We generate a single output per test example. The inputs are the human rewritings and outputs from our two-stage models, and a subset of one-stage models. Each example is evaluated by two annotators. See Appendix D for more.

They evaluate fluency, coherence, logic, and plot preservation (plot-pres) on 1-5 scales. Fluency is a measure of how fluent and readable a text is. Coherence is how well individual sentences fit together (Barzilay and Lapata 2008). Logic is the plausibility of described events. Plot-pres is how well reordered text preserves the plot of the original. This includes details about characters, events, and interactions between them, encompassing its semantic and temporal aspects.

We also conduct an interestingness (interest) study on human rewritings and outputs from our BART-2S and T5-2S models. Each reordered story’s interestingness w.r.t. suspense and time flow compared to the original are evaluated from 1-5 by two annotators. We ask the following: “On a scale of 1-5, with 1 being most decrease in interestingness and 3 being same level of interestingness and 5 being most increase in interestingness, how interesting is the suspense and flow of time in the story S, compared to the original story O? How exciting did you find the story as you read through it?”

5 Results and Analysis

We present evaluation results of our 2S and subset of 1S models on testSup compared to human rewritings and original stories. Tables 2 and 3 contain human evaluation results, and Table 4 automatic evaluation results. Correlations between automatic and human metrics are in Table 5. Table 6 contains qualitative examples, with more in Appendix E.

5.1 Analysis of Human Evaluation Results

We begin by analyzing human evaluation performance through results in Tables 2 and 3.

7Nucleus sampling did not work as well for BART and T5.

8See Appendix C for more finetuning/generation details.
We now analyze the automatic evaluation performance of the different methods in Table 4.

### 5.2 Analysis of Automatic Evaluation Results

#### BERTScore, BLEU, METEOR:

We see from Table 5 that these reference-based metrics correlate quite well with human eval metrics, particularly plot-pres. T5-d-2S performs best followed by BART-d-2S. Similar to the human evaluation, 2S models outperform their 1S variants, and GPT-2 models perform worst overall. Denoise outperforms reorder variants and generate more similar text, on avg, to human references.

#### Fluency, Coherence, Logic:

Original stories are the highest for all three metrics with human rewritings second for coherence and logic, beating the models by a noticeable degree. BART-d-2S and T5-r-2S are generally the best-performing models here. BART-d-2S slightly outperforms human rewritings on fluency, with T5-r-2S closely behind, demonstrating that these models are quite fluent. These models also outdo their 1S variants. GPT-2 models perform worst on all metrics.

#### Plot-pres:

We see that human rewritings best preserve the plot of the original stories. T5-d-2S is the best performing model on plot-pres, followed by BART-r-2S and T5-r-2S. GPT-2 models perform the worst at preserving the plot of the original stories (which we show qualitatively in §5.3).

#### Interestingness:

Human rewritings score highest on interestingness. Humans rewrite the text in more creative ways, whereas BART and T5 models are more conservative (see §5.2 TOF and §5.3). Narrative reorderings for all methods are more interesting, on average, than original stories. NAREOR can indeed be used to generate more interesting story variations.

### 5.3 Qualitative Analysis

From Table 6, we see that humans modify text heavily to suit the reorderings and are sometimes quite creative, e.g. phrasing Fred as having grown accustomed to the bird being his alarm clock (ex. 2). Humans successfully handle necessary coreferences, tenses, time expressions (timexes), etc.

GPT-2 modifies text quite heavily but suffers from incorrect coreference while introducing spurious tokens, repetition, or hallucinations. For ex. 2, GPT2-r changes the plot greatly, stating Fred woke him up for work and This was because he liked Fred (likely due to poor coreference), and hallucinating This bird, however, did not like Fred. For ex. 4, it repeats Joey’s excitement many times, while hallucinating a roller coaster that was absent in the original story.

BART and T5 models are more conservative, but their edits are important and effective. They handle coreference, tense, and timexes quite well. These pinpointed and critical edits are required to maintain plot. For ex. 1, they modify He told him that he was lost to Jimmy told the policeman that he was lost given that sentence is now at the beginning. BART-d impressively modifies tense by converting Soon a policeman came by and asked if he was lost to The policeman had come by and asked if he had been lost. For ex. 2, T5-d converts enjoyed to had enjoyed since the bird no longer singing is now prior information, and adds the timex After a while to the beginning of the last output sentence. BART-r successfully changes Fred began to like the bird to He had begun to like the bird. For ex. 3, BART-d inserts the timex Earlier at the beginning of the second output sentence, correctly and unambiguously conveying its underlying temporality w.r.t. the first. BART-d correctly changes saw a turtle to had seen a turtle, while BART-r does so for stepped to had stepped. For ex. 4, BART and T5 models all resolve the Disneyland ellipsis by converting Joey had a great time to Joey had a great time at Disneyland, while GPT2-d cannot.

However, the BART and T5 models are imperfect. For ex. 1, BART-r halluciates lost his wallet (original story does not involve a wallet), T5-d inserts an incorrect timex of Soon after at the beginning of the second output sentence, and T5-r hallucinates asked if he had a soda (this is not asked in the original story). For ex. 2, BART-r incorrectly converts the bird no longer sang to Fred no longer sang, likely due to coreference difficulties. For ex. 3, T5-r does not convert Suddenly to Earlier like BART-d, giving a false interpretation that Eric slipped after his rescuer’s arrival. BART-r does not mislead with Suddenly, but is ambiguous and has no timex.

---

Although these metrics slightly decrease for reordered stories, we note that NAREOR’s main purpose is for more interesting tellings of the same story which we do achieve (see Table 3).
Table 5: Pearson and Spearman correlations between automatic and human evaluation metrics, with p-values in brackets. TOF models excluded as they are mainly for validation. Bold corresponds to highest correlation per human evaluation metric.

Table 4: Average automatic evaluation results on testSup (values multiplied by 100). Bold corresponds to best performance per metric, and underline second-best (excluding the TOF metrics which are mainly for validation).

5.4 Overall Takeaways

Humans modify text greatly while successfully performing NAREOR. BART and T5 models perform decently with minimal but effective edits. GPT-2 models tend to repeat, hallucinate, and reduce text quality and plot preservation. Based on human (§5.1) and automatic (§5.2) evaluation, BART-d-2S and T5-d-2S are the best models overall. BART-d-2S outdoes its reorder variant, possibly due to BART’s pretraining as a denoising autoencoder, closer to our denoise training method. For T5, both methods perform quite well and show potential. However, T5-d outperforms on plot-pres (Table 2), interest (Table 3), and automatic metrics (Table 4). The denoise training method appears to be slightly more effective, possibly because it is partially inspired by how humans perform NAREOR (see §3.1). These are the first two task-specific training methods for NAREOR which we propose ourselves, each approaching the task differently (see §3.1). 2S models also mostly outperform 1S ones, demonstrating that second stage finetuning improves upon the first.

BART and T5 models are quite effective, excelling at fluency, but have further room for improvement in coherence, logic, plot-pres, and interest. §5.3 shows they still suffer from several issues. Their conservative tendency may limit their NAREOR ability compared to humans. Overall, these models serve as strong initial baselines for NAREOR while under-scoring the task’s difficulty and potential for exploration.

6 Applications of NAREOR

Sentence ordering involves reconstructing original sentence order of an unordered sentence set (Barzilay and Lapata 2008). NAREORC’s reordered stories could serve as a challenge set for sentence reordering models due to their non-linear narrative structure underrepresented in corpora. We use the implementation of (Prabhumoye, Salakhutdinov, and Black 2020) to train i) $M_{ext}$, an external model on the SIS corpus (Huang et al. 2016), ii) $M_{iid}$, an in-domain model on first 20% of ROCStories’ train split. We test each on i) Control set $\{s_i\}_{i=1}^{n}$, input stories from testSup, ii) Challenge set $\{s'_i\}_{i=1}^{n}$, reordered stories from testSup. Table 7 shows drastic drops across metrics (higher is better - see Prabhumoye, Salakhutdinov, and Black (2020)) for both $M_{ext}$ and $M_{iid}$ from control to challenge set, confirming our hypothesis. Systems with ability to manipulate narrative variables like order could be important for automating pedagogical setups, especially for fine-grained language skills such as argumentation in essay writing. As Wingate (2012) explains, tutor understanding is found deficient and methods of feedback for students are inconsistent or vague. Language in school texts follows a characteristic register, which often differs from registers students handle in everyday conversation (Schleppegrell 2001). Models (e.g. NAREOR ones) which can control elements of register, e.g narrative order, can be used to tailor such content to intended settings and bridge this gap. Systems that can generate event timelines for clinical narratives, e.g. admission notes and physical reports, are important for applications like medical document summarization (Bramsen et al. 2006; Reichert et al. 2010) and clinical decision making (Demner-Fushman, Chapman, and McDonald 2009). Raghavan et al. (2014) demonstrate that cross-narrative temporal ordering of medical events is vital to generating a comprehensive timeline over a patient’s history. Aligning multiple medical event sequences using coreference information and temporal relations has a large impact on their presentation and effectiveness. Our NAREOR models may be effective here and improve upon existing systems.

7 Related Work

There exists work on the sentence ordering task discussed in §6 — e.g., Chen, Qiu, and Huang (2016) learn pairwise orderings of sentences using a ranking model. Unlike sen-
Fred didn’t hear the bird singing this morning which is unusual. The bird had been waking him up every single day at 6 AM for some time. He had grown accustomed to the bird being his alarm clock. Now he’s worried that something might have happened to the poor bird. He almost considers the bird a friend at this point.

Kendall

Someone walking on the shore ran over to rescue Eric. Eric had been walking around a pond while at a local park. Trying to catch a turtle in the pond, Eric stepped into the water. Eric did so because he had seen a turtle poached on a rock few feet offshore.

Jimmy told a policeman that he was lost. A policeman came by and asked if he had a soda. Before he knew it, he was in a strange area. He had wandered around town looking for a soda. He was scared of strangers and didn’t want to ask anyone.

Joey had a great time! He was so excited to meet Mickey Mouse. He finally said hi to Mickey and it was great! He had gone to Disneyland with his family. When he met Mickey Mouse he was so excited he couldn’t speak! (...)

One day, Fred’s bird began to sing. Every morning, Fred woke up for work at 6 AM. This was because he liked Fred. He had hoped nothing bad had happened to Fred. This bird, however, did not like Fred any longer. Every morning, he woke Fred up for work at 6 AM. He enjoyed the sound of its singing. He hoped nothing bad had happened to him. He had begun to like the bird.

We proposed the NAREOR task and introduced a dataset, NAREORC, with task-specific training methods and evaluation metrics, and experimented with T5, BART, and GPT-2. Extensive evaluation and analysis showed that our models are effective but can be further improved, and that NAREOR is challenging with further exploration potential. We showed that NAREOR can create interesting story variations and challenge sets for tasks like sentence ordering.

Future directions include exploring training ideas better emulating human rewrites. NAREOR can be explored as document-level paraphrasing for data augmentation, adversarial sets for more temporal tasks, and applications for education/medicine (see §6). We also hope our work drives inquiry into harder task variations of NAREOR (e.g. sub-sentential).

### Table 6: Qualitative testSup examples. Target perms in brackets along original stories. d & r refer to denoise & reorder.

<table>
<thead>
<tr>
<th>Model</th>
<th>TextSet</th>
<th>SentAcc</th>
<th>Rouge-S</th>
<th>LCS</th>
<th>Kendall τ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mext</td>
<td>Control</td>
<td>70.35</td>
<td>48</td>
<td>59.1</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>Challenge</td>
<td>52.4</td>
<td>24.7</td>
<td>29.7</td>
<td>0.12</td>
</tr>
<tr>
<td>Midd</td>
<td>Control</td>
<td>66.4</td>
<td>85.3</td>
<td>84.8</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Challenge</td>
<td>21.9</td>
<td>49.6</td>
<td>58</td>
<td>0.03</td>
</tr>
</tbody>
</table>

### Table 7: Sentence ordering on control vs. challenge sets.
References


