Carnegie Mellon University
Language Technologies Institute


# #3174
# NAREOR: The Narrative Reordering Problem

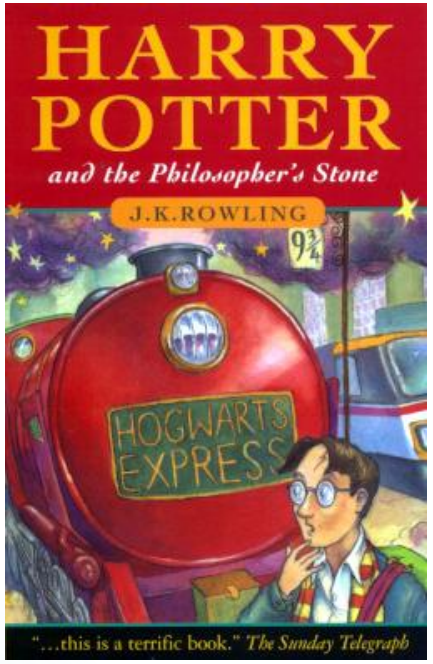
Varun Gangal*[1], Steven Y. Feng*[1], Malihe Alikhani[2], Teruko Mitamura[1], Eduard Hovy[1]

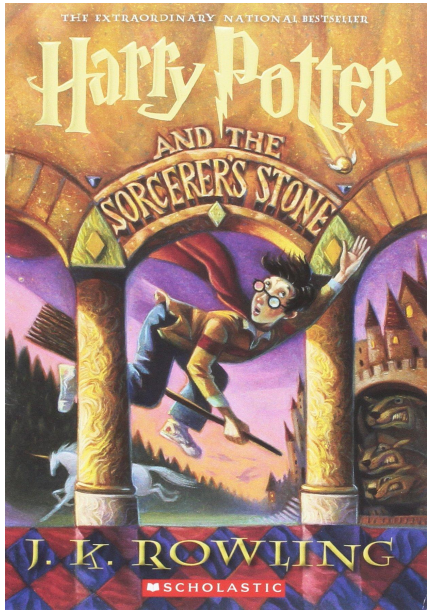
[1]Language Technologies Institute, Carnegie Mellon University

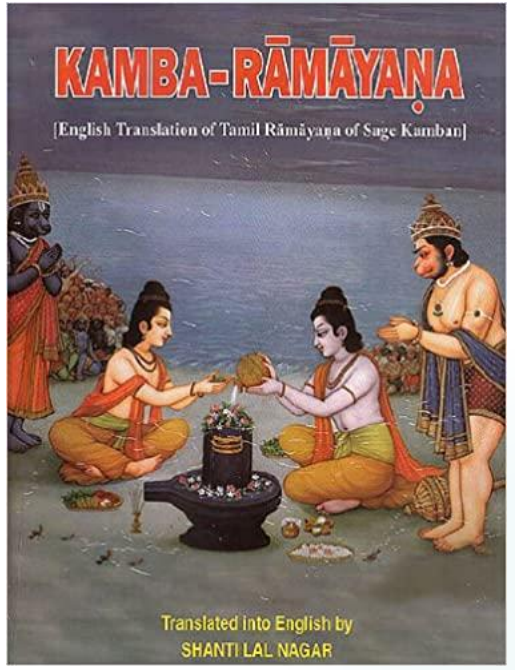
[2]University Of Pittsburgh

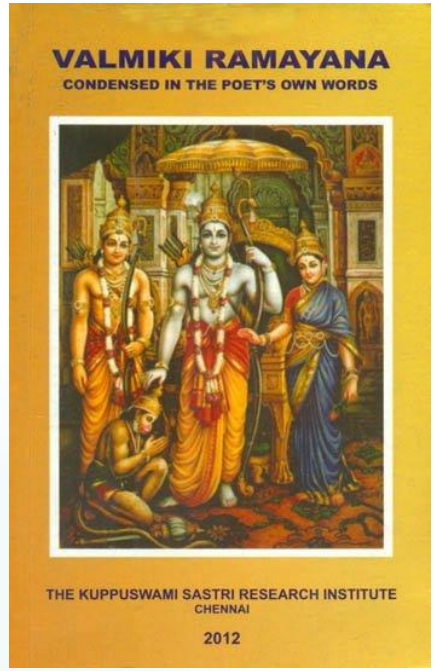
## AAAI '22

https://arxiv.org/abs/2104.06669

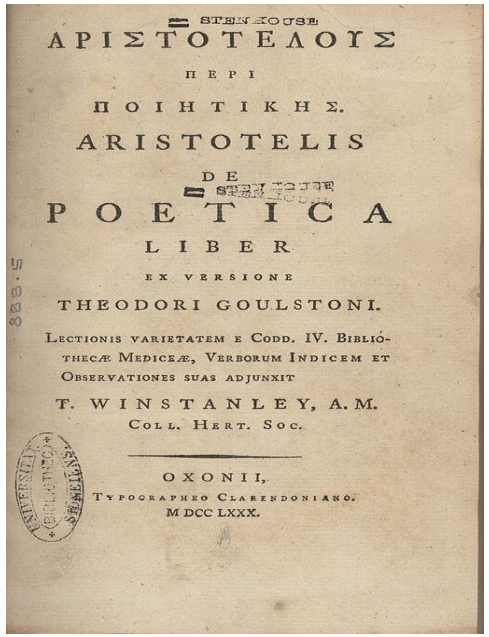



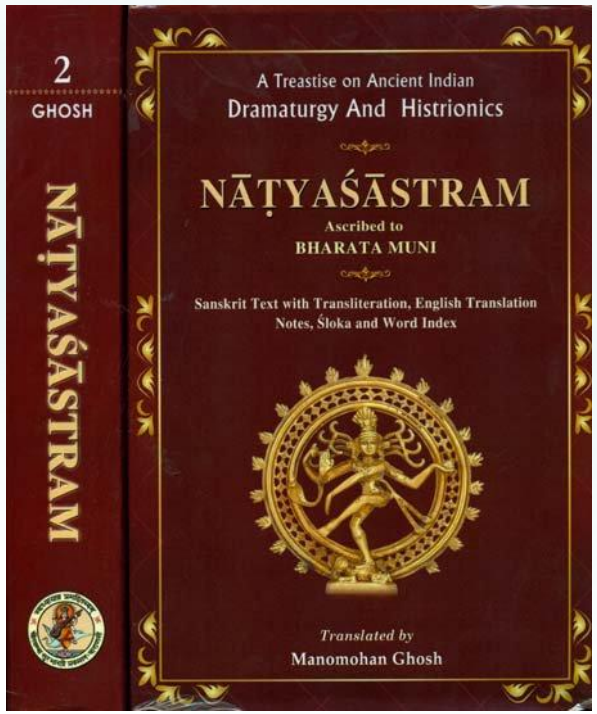



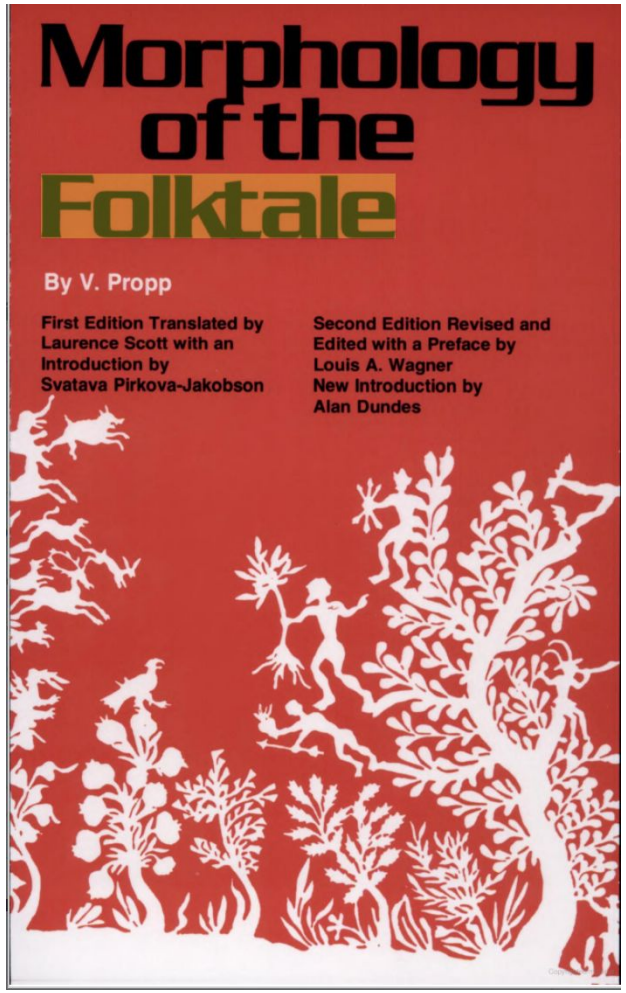

- Emergence of Storytelling and Multiple Versions a.k.a "Narratives"
  - Listener remembers only underlying story abstraction (Schank et al, '72)
  - Tailors to new audience ; Impromptu improvisation
  - "Favourite elements" e.g characters → Caesar vs Brutus, Harry Potter vs Snape
  - Thus we have 4 Gospels (Matthew-Luke-Mark-John) and 300 Ramayanas

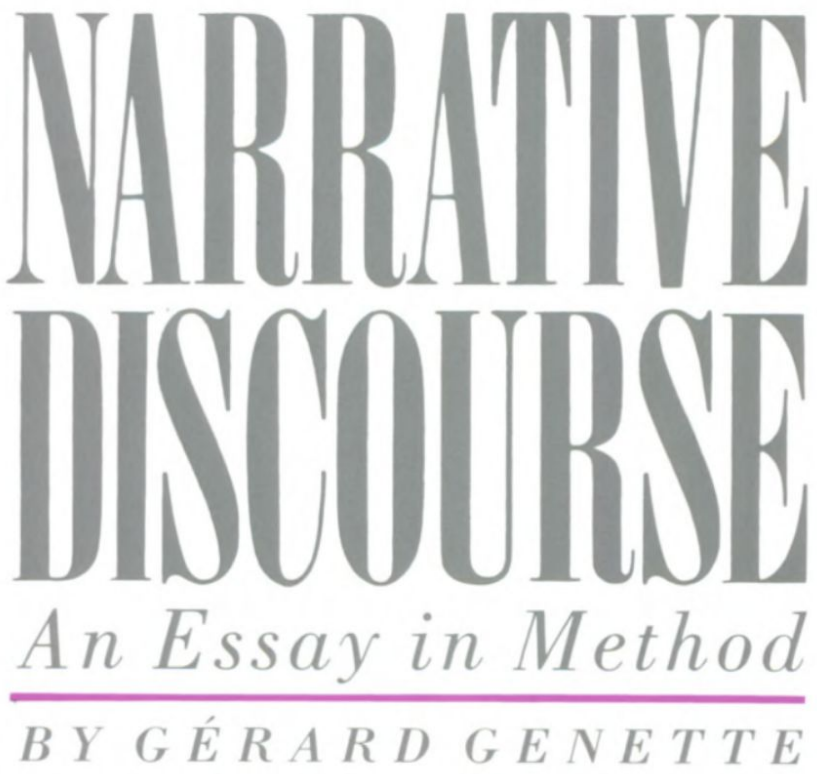
# Narrative Theory & Aspects of Narrative Style

- Formal analysis of Story and Narrative → Aristotle's *Poetics*
- *Story* → **What** underlying events were Vs *Narrative* → **How** they are told in text
- Narrator can make different choices about how to present the narrative
- *Story Order* = Chronological order of events in actual, underlying story
- Many Aspects of Narrative Style a.k.a Elements of Narrativity (Genette, 1983)
  - First Person vs Third Person Narrator
  - Omnipresent Narrator vs Character-as-Narrator
  - Temporality or *Narrative Order* = Order of presenting events in the text


ΑΡΙΣΤΟΤΕΛΟΥΣ
ΠΕΡΙ
ΠΟΙΗΤΙΚΗΣ.
ARISTOTELIS
DE
POETICA
LIBER
EX VERSIONE
THEODORI GOULSTONI.
Lectionis varietatem e Codd. IV. Bibliothecæ Mediceæ, Verborum Indicem et Observationes suas adjunxit
T. WINSTANLEY, A.M.
Coll. Hert. Soc.
OXONII,
Typographeo Clarendoniano.
M DCC LXXX.


2
GHOSH
NĀṬYAŚĀSTRAM
A Treastise on Ancient Indian
Dramaturgy And Histrionics
NĀṬYAŚĀSTRAM
Ascribed to
BHARATA MUNI
Sanskrit Text with Transliteration, English Translation
Notes, Śloka and Word Index
Translated by
Manomohan Ghosh


Morphology of the Folktale
By V. Propp
First Edition Translated by Laurence Scott with an Introduction by Svatava Pirkova-Jakobson
Second Edition Revised and Edited with a Preface by Louis A. Wagner
New Introduction by Alan Dundes


NARRATIVE DISCOURSE
An Essay in Method
BY GÉRARD GENETTE

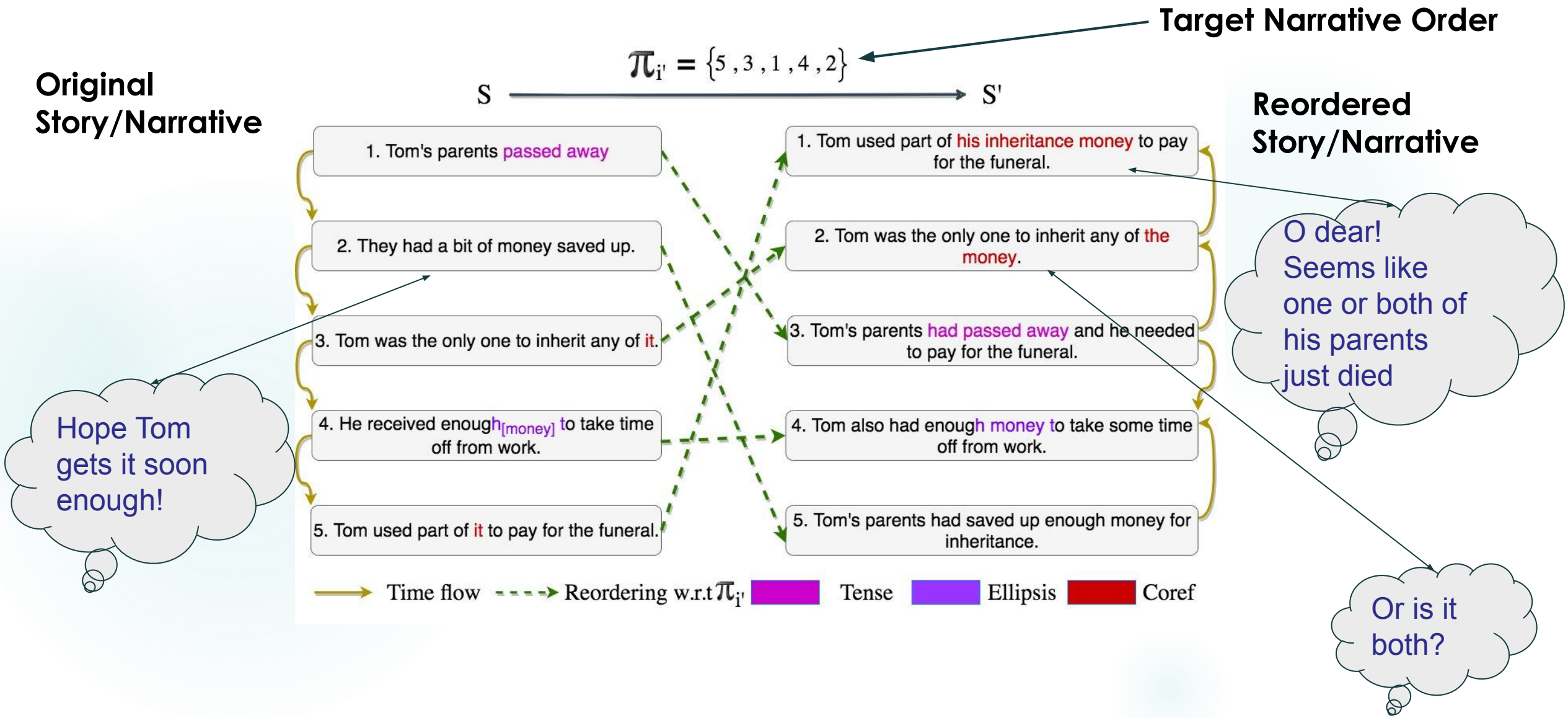
# *NAREOR*: Motivation and Introduction

- Many implicit inferences can be drawn at different points in a text depending on how it is structured that can critically impact the text's evolving interpretation and meaning in the reader's mind.
- Reordering a narrative → Alters these temporal, causal, event-based, and other inferences readers can draw from it → *Sequence* and *ease* of interpretation, *interestingness*, *suspense*, *counterfactual* thought process
- We propose the task of Narrative Reordering a.k.a *NAREOR*:
  - Rewrite a given linear narrative **S** → A different, target narrative order $\mathbf{\Pi}_i$,
  - Preserve plot and all other elements of the underlying story

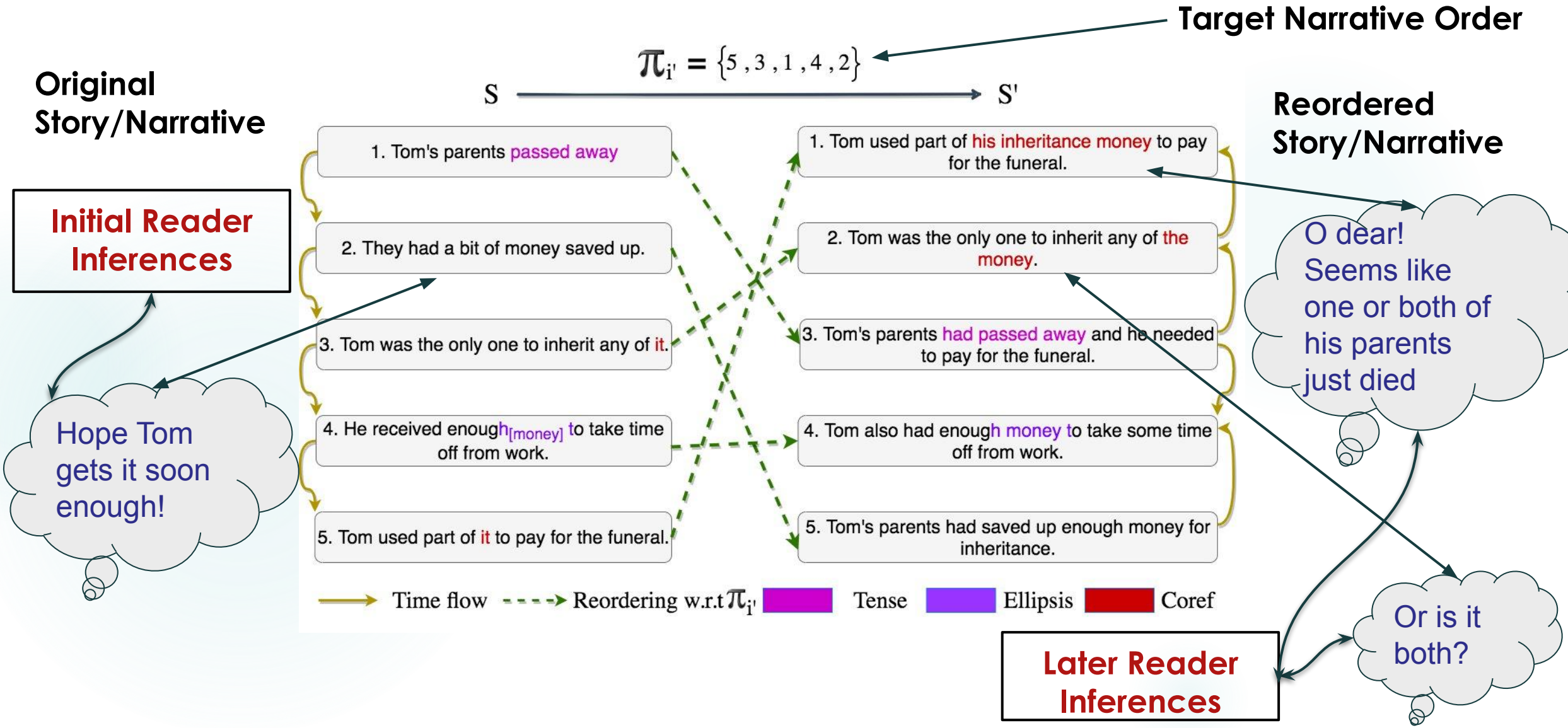
# *NAREOR*: Motivation and Introduction

- Many narrative orders[10,11]:
  - *Linear*: narrates events in chronological sequence
  - *In media res*: starts with events in the middle, goes back to the start, then to the end
- Changing to more "interesting" orders is prevalent in cinema and entertainment
  - The Imitation Game, The Iliad, Citizen Kane (In medias res - "in the middle")
  - Memento (Retrograde – reverse of linear)
  - Naked Lunch (Syllepsis – lacking chronological logic)
  - 500 Days of Summer (Syllepsis)
  - Many, many more examples…

# NAREOR: Illustrated Example



# NAREOR: Illustrated Example



# *NAREOR*: Challenges

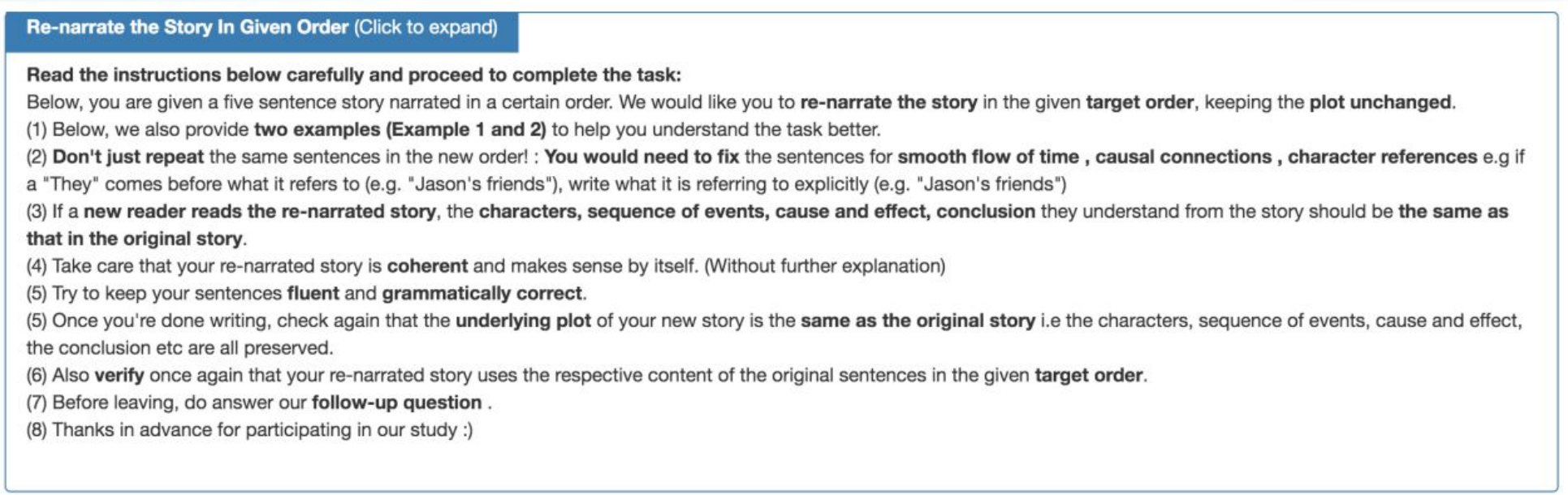
- Critical and pinpointed edits to *maintain plot* and ensure valid story
- Rewritten text must be adjusted to handle coreference, tense, timexes , ellipsis, cross-sentence event arguments ….
- Challenging controllable text generation task since:
  1. Control variable (target order) → Many-valued (n!-1) and Complex
  2. Invariant of faithfulness (plot preservation) ↔ Strong understanding of story event order, characters, interactions, etc.
  3. Extra-sentential and discourse sensitive: Generating a full, coherent story and learning several types of discourse dependencies

# *NAREOR*: Contributions & Outline

- Curate a dataset, *NAREORC*, with human rewritings of stories within ROCStories in non-linear orders, and conduct a detailed analysis of it
- Propose novel task-specific training methods with suitable evaluation metrics
- Experiments on *NAREORC* using SOTA generators (GPT2, BART, T5)
- Thorough Automatic and Human evaluation; Qualitative Analysis
- Discussion on Applications; mini-experiments for 2 downstream Applications

# *NAREORC* Dataset

- ROCStories Corpus (Mostafazadeh, '17): ~98K 5-sentence stories
- Assigning $\mathbf{\Pi}_i$, → Sample 3 non-identity permutations → Pick lowest Kendall one
- Annotations for 1K stories → 600 trainSup, 200 devSup, 200 testSup
- 95161 trainUnsup, 1671 devUnSup, 1671 testUnsup → Retained for unsupervised learning



**Re-narrate the Story In Given Order** (Click to expand)

**Read the instructions below carefully and proceed to complete the task:**
Below, you are given a five sentence story narrated in a certain order. We would like you to **re-narrate the story** in the given **target order**, keeping the **plot unchanged**.
(1) Below, we also provide **two examples (Example 1 and 2)** to help you understand the task better.
(2) **Don't just repeat** the same sentences in the new order! : **You would need to fix** the sentences for **smooth flow of time , causal connections , character references** e.g if a "They" comes before what it refers to (e.g. "Jason's friends"), write what it is referring to explicitly (e.g. "Jason's friends")
(3) If a **new reader reads the re-narrated story**, the **characters, sequence of events, cause and effect, conclusion** they understand from the story should be **the same as that in the original story**.
(4) Take care that your re-narrated story is **coherent** and makes sense by itself. (Without further explanation)
(5) Try to keep your sentences **fluent** and **grammatically correct**.
(5) Once you're done writing, check again that the **underlying plot** of your new story is the **same as the original story** i.e the characters, sequence of events, cause and effect, the conclusion etc are all preserved.
(6) Also **verify** once again that your re-narrated story uses the respective content of the original sentences in the given **target order**.
(7) Before leaving, do answer our **follow-up question** .
(8) Thanks in advance for participating in our study :)

# *NAREORC* Dataset

**Re-narrate the Story In Given Order** (Click to expand)

**Read the instructions below carefully and proceed to complete the task:**

Below, you are given a five sentence story narrated in a certain order. We would like you to **re-narrate the story** in the given **target order**, keeping the **plot unchanged**.

(1) Below, we also provide **two examples (Example 1 and 2)** to help you understand the task better.

(2) **Don't just repeat** the same sentences in the new order! : **You would need to fix** the sentences for **smooth flow of time , causal connections , character references** e.g if a "They" comes before what it refers to (e.g. "Jason's friends"), write what it is referring to explicitly (e.g. "Jason's friends")

(3) If a **new reader reads the re-narrated story**, the **characters, sequence of events, cause and effect, conclusion** they understand from the story should be **the same as that in the original story**.

(4) Take care that your re-narrated story is **coherent** and makes sense by itself. (Without further explanation)

(5) Try to keep your sentences **fluent** and **grammatically correct**.

(5) Once you're done writing, check again that the **underlying plot** of your new story is the **same as the original story** i.e the characters, sequence of events, cause and effect, the conclusion etc are all preserved.

(6) Also **verify** once again that your re-narrated story uses the respective content of the original sentences in the given **target order**.

(7) Before leaving, do answer our **follow-up question** .

(8) Thanks in advance for participating in our study :)

# *NAREORC*: Examples & Change Type Stats

| Change Type | Story Examples with Changes Highlighted |
|---|---|
| Ellipsis<br>(Sent: 5.7%)<br>(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.* ; $\pi_{i'}$: $\{1, 5, 4, 2, 3\}$<br>**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<br>(Sent: 19.1%)<br>(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. ; $\pi_{i'}$: $\{2, 3, 5, 1, 4\}$<br>**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<br>(Sent: 34.0%)<br>(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. ; $\pi_{i'}$: $\{4, 2, 5, 1, 3\}$<br>**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<br>(Sent: 20.7%)<br>(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.* ; $\pi_{i'}$: $\{5, 4, 2, 1, 3\}$<br>**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. |

# *NAREOR*: Models

- Base Architectures → GPT2, BART, T5
  - GPT2 → Sequential Language model
    (No Separate Encoder for Input)
  - BART, T5 → Sequence2Sequence model
    (Encoder for Input & Decoder for Output)

- Can't simply do prompt-continuation finetuning → Need a way to control for $\mathbf{\Pi}_i$,

# *NAREOR*: Models

- Base Architectures → GPT2, BART, T5
  - GPT2 → Sequential Language model

    (No Separate Encoder for Input)
  - BART, T5 → Sequence2Sequence model

    (Encoder for Input & Decoder for Output)

- Can't simply do P(S' | S) finetuning → Need a way to input and control for $\mathbf{\Pi}_i$

# Models - NAR-reorder a.k.a NAR-r

- Encode $\mathbf{\Pi}_i$, using special lettered tags

*<a> Since I had front seat tickets, I was able to directly see <X1>. <b> <X1> tried to reach out with <X1> <X2>. <c> I grabbed <X2> and <X1> pulled me on stage. <d> <X1> began to sing. <e> The concert had started. <sep> <e> <d> <a> <b> <c> <X1> The music artist <X2> her hand <st>*

- Separate out coreference chains
- Train in inverse direction $\mathbf{S'_{naive} + \Pi_i^{-1} \rightarrow S}$ for Stage 1 Unsupervised Training
- Train in typical direction $\mathbf{S + \Pi_i \rightarrow S'}$ for Stage 2 Supervised Training

# Models - NAR-denoise a.k.a NAR-d

- Naively reorder S → $S'_{naive}$ as per $\Pi_i$, first
- Model has to simply rewrite sentences to preserve plot → Relieved of responsibility of reordering
- For Stage 1 Unsupervised Training, we lack actual target narratives S'
  - How do we get over this? Use a *Denoising* setup!
  - Create pseudo-target narratives S' by randomly deleting and swapping tokens
  - Train model to reconstruct original story narrative S as target
- Stage 2 Supevised Training: $S'_{naive}$ → S'

# Metrics

- Reference Matching: BLEU, METEOR and BERTScore
- How do we know if rewritten target narratives actually follow specified order?
  - Desired: A “sanity check” metric
    - Target Order Fidelity a.k.a TOF
    - Let $\mathbf{\Pi}_{i}$(K) = J (Kth sentence in target narrative comes from Jth in original)
    - Rewritten sentence at position K in target narrative should be largely similar to original sentence at position J, barring few edits to adjust to altered narrative order → Eval using BLEU, METEOR etc

# Automatic Evaluation Results

| *Method\Metric* | BERTScore | BLEU | METEOR | TOF-BERTScore | TOF-METEOR |
|---|---|---|---|---|---|
| **Human rewritings** | N/A | N/A | N/A | 66.85 | 56.79 |
| **GPT2-d-2S** | 60.75 | 37.01 | 45.20 | 79.23 | 74.23 |
| **GPT2-r-2S** | 58.03 | 32.57 | 40.85 | 73.04 | 63.00 |
| **BART-d-1S** | 67.14 | 44.73 | 49.88 | 95.61 | 93.43 |
| **BART-d-2S** | 67.93 | 46.03 | 50.54 | 93.55 | 90.81 |
| **BART-r-2S** | 67.16 | 44.63 | 49.16 | 91.32 | 86.43 |
| **T5-d-2S** | **67.99** | **46.95** | **51.12** | 94.20 | 91.83 |
| **T5-r-1S** | 66.24 | 43.40 | 48.20 | 89.85 | 84.26 |
| **T5-r-2S** | 66.62 | 44.30 | 49.00 | 91.61 | 86.16 |

- **T5-d-2S** best model across the board on matching the reference
- Two stage models better than one stage counterparts
- **BART-*** and **T5-*** models have high TOF, even higher than Human → This is in part an effect of them being more conservative with edits (as we'll see), and illustrates why TOF should only be used as a sanity check metric and not be optimized towards

# Human Evaluation Results

| *Method\Metric* | Fluency | Coherence | Logic | Plot-pres |
|---|---|---|---|---|
| **Original stories** | 4.209 | 4.0 | 3.851 | N/A |
| **Human rewritings** | 3.797 | 3.723 | 3.784 | 3.972 |
| **GPT2-d-2S** | 3.635 | 3.399 | 3.399 | 3.708 |
| **GPT2-r-2S** | 3.595 | 3.378 | 3.291 | 3.375 |
| **BART-d-1S** | 3.628 | 3.412 | 3.318 | 3.847 |
| **BART-d-2S** | **3.818** | 3.507 | 3.493 | 3.722 |
| **BART-r-2S** | 3.757 | 3.439 | 3.493 | 3.861 |
| **T5-d-2S** | 3.764 | 3.419 | 3.5 | **3.889** |
| **T5-r-1S** | 3.655 | 3.378 | 3.486 | 3.847 |
| **T5-r-2S** | 3.784 | **3.595** | **3.520** | 3.861 |

- **T5-r-2S** seems to be the best model across the board.
- **T5-d-2S** is the most plot-preserving
- Overall, all **BART-*** and **T5-*** variants are mutually comparable and reasonable on all four metrics, and comprehensively outperform **GPT2-*** variants, while narrowly trailing **Human.**
- Differences b/w BART+T5 and GPT2 → The first two are encoder-decoder, while the latter is just sequential

# Qualitative Analysis - I

| | |
|---|---|
| Original (ex. 1) | Jimmy wandered around town looking for a soda. Before he knew it, he was in a strange area. He was scared of strangers and didn't want to ask anyone. Soon a policeman came by and asked if he was lost. He told him that he was lost. {5,4,2,1,3} |
| BART-d | **Jimmy told a policeman** that he was lost. The policeman **had come** by and asked if he **had been** lost. Before he knew it, he was in a strange area. Jimmy had wandered around town looking for a soda. He was scared of strangers and didn't want to ask anyone. |
| BART-r | **Jimmy told the policeman** that he was lost. The policeman came by and asked if he had lost **his wallet**. Before Jimmy 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. |
| T5-d | **Jimmy told a policeman** that he was lost. **Soon after**, a policeman came by and asked if he was lost. 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. |
| T5-r | **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. |

- All models able to resolve the two "*he*"s correctly to their named mention as the sentence moves to first position → *He told him that he was lost* to *Jimmy told a/the policeman that he was lost*
- BART-d adept at tense changes → Introduces "had come", "had been"
- Hallucinations still a significant concern → BART-r dreams up "his wallet", T5-r dreams up "if he had a soda". This alters the underlying plot

# Qualitative Analysis - II

| | |
|---|---|
| Original (ex. 2) | Every morning, a bird woke Fred up for work at 6 AM. Fred began to like the bird. And he enjoyed the sound of its singing. One day, the bird no longer sang. He hoped nothing bad had happened to the bird. {4,1,3,5,2} |
| Human | 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 quite 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. |
| GPT2-r | One day, Fred's bird began to sing. Every morning, **Fred woke him up** for work at 6 AM. **This was because he liked Fred**. He hoped nothing bad had happened to Fred. **This bird, however, did not like Fred**. |
| BART-r | **Fred** no longer sang. Every morning, he woke Fred up for work at 6 AM. He enjoyed the sound of his singing. He hoped nothing bad had happened to him. He **had begun** to like the bird. |
| T5-d | One day, the bird no longer sang. Every morning, a bird woke Fred up for work at 6 AM. He **had enjoyed** the sound of its singing. He had hoped nothing bad had happened to the bird. **After a while**, Fred began to like the bird. |

- GPT2-* seems more aggressive with rewriting than T5-* and BART-* - but hallucinates (the bird *liked Fred*, and then *did not like* him) & misattributes roles (*Fred woke up the bird for work*) considerably more
- T5-d: *enjoyed* → *had enjoyed* ✔ since "*bird no longer sang*" is now prior info
- T5-d: Timex "*After a while*" to beginning of last output sentence ✔ Valid alternative to doing *began* → *had begun* (Choice taken by BART-r)

# Encoder-Decoder vs Simple Sequential - Why the former might be better

- **T5-r-2S** seems to be the best model across the board.
- **T5-d-2S** is the most plot-preserving
- Overall, all **BART-*** and **T5-*** variants are mutually comparable and reasonable on all four metrics, and comprehensively outperform **GPT2-*** variants, while narrowly trailing **Human.**
- Differences b/w BART+T5 and GPT2 → The first two are encoder-decoder, while the latter is just sequential

# Qualitative Analysis - Takeaways

- BART and T5 make minimal but precise edits, especially for specific grammatical phenomena like tense, ellipsis etc.
- Human rewritings  are much more aggressive than BART and T5, though free of hallucinations.
- GPT2 is less conservative than BART and T5, but hallucinates wildly.
- GPT2 also suffers from repetition; overall poor plot preservation compared to BART and T5.

# Applications

- We investigate 2 applications of *NAREOR*:
  - Generation of more "*interesting*" variants of stories
  - Serving as adversarial sets for temporal/event-based tasks
- Other exciting possibilities, such as
  - Pedagogical setups related to language skills like essay writing
  - Applications to medicine involving clinical narratives (Reduce suspense)
- Document/story-level data augmentation for multi-sentence tasks

# Application I : Interestingness

- How interesting are stories rewritten as per target narrative orders compared to original stories? We ask annotators to answer on 1-5 → 3 = equivalent
- Both **Human** and all model variants **BART-*** and **T5-*** models generate more interesting stories than the original.
- **T5-d** and **BART-r** are considerably better than their architectural siblings.

| *Method:* | Human | BART-d | BART-r | T5-d | T5-r |
|---|---|---|---|---|---|
| **Interest** | **3.75** | 3.367 | 3.483 | 3.533 | 3.3 |

Average interestingness results on testSup, rated from 1-5 (3 represents equal to original story). Models are 2S versions. Bold corresponds to best performance, and underline second-best.

# Application II :Challenge Set for Temporal Tasks?

- Drastic Performance Drop from Control → Challenge
  - For both externally trained model $\mathbf{M_{ext}}$ and in-domain trained $\mathbf{M_{iid}}$
  - Across all 4 metrics
- *NAREOR* reordered stories indeed a strong Challenge set for Sentence ordering!

| *Model* | *TestSet* | SentAcc | Rouge-S | LCS | Kendall $\tau$ |
|---|---|---|---|---|---|
| $M_{ext}$ | Control | 76.35 | 48 | 59.1 | 0.57 |
| | Challenge | 52.4 | 24.7 | 29.7 | 0.12 |
| $M_{iid}$ | Control | 66.4 | 85.3 | 84.8 | 0.75 |
| | Challenge | 21.9 | 49.6 | 58 | 0.03 |

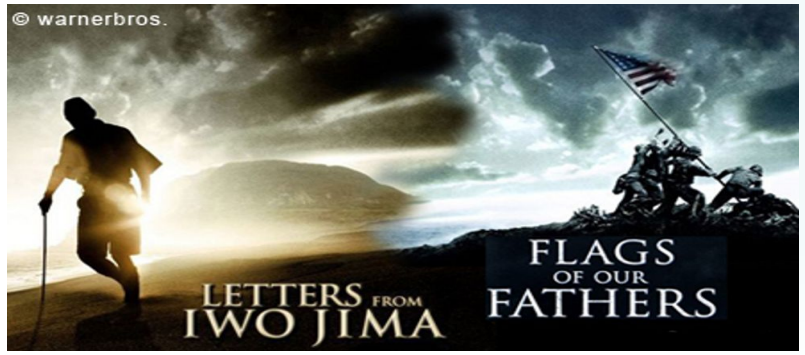
Sentence ordering on control vs. challenge sets.

# Conclusions

- Proposed *NAREOR* and Curated *NAREORC*.
- Proposed novel, task-specific training atop SOTA generators, and devise suitable evaluation metrics.
- Though our models perform decently on both auto and human evaluation, *NAREOR* is a challenging task with potential for further exploration.

# Future Work

- Exploring other aspects of narrative style similarly
  - First person vs Third person Narrator, Omnipresent vs Character-as Narrator
  - Character and Story Element Focus




- Learning to generate draw on-the-fly inferences to estimate interestingness.
- Using *NAREOR* for document-level data augmentation.

# Check Out Other Works From Our Stable!

- On Data Augmentation:
  - Genaug: Data augmentation for finetuning text generators (EMNLP'20 WS)
  - Keep calm and switch on! preserving sentiment and fluency in semantic text exchange (EMNLP '19)
  - A survey of data augmentation approaches for NLP (Findings of ACL '21)
  - Improving Automated Evaluation of Dialog via Diverse Reference Augmentation (Findings of ACL '21)
  - NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation (Preprint)
- On Generative Common Reasoning:
  - SAPPHIRE: Approaches for Enhanced Concept-to-Text Generation (INLG '21)
  - Retrieve, Caption, Generate: Visual Grounding for Enhancing Commonsense in Text Generation Models (Also at AAAI'22!)

# That's all Folks!

# Qualitative Analysis - III

| | |
|---|---|
| Original (ex. 4) | Joey went to Disneyland with his family. He was so excited to meet Mickey Mouse. When he met Mickey Mouse he was so excited he couldn't speak! Joey finally said hi to Mickey and it was great! Joey had a great time. {5,2,4,1,3} |
| GPT2-d | Joey had a great time. He was so excited to meet Mickey Mouse. He had 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 |
| GPT2-r | Joey had a great time on the **roller coaster** at Disneyland. **He was so excited** to meet Mickey Mouse he couldn't speak! (...) **He was so excited** that he went to the park with his family. **He was so excited** when he met Mickey Mouse and **he was so excited**. |
| BART-r | Joey had a great time **at Disneyland**. 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 he was excited he couldn't speak! |
| T5-d | Joey had a great time **at Disneyland**. He was so excited to meet Mickey Mouse. He had 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! |

- 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.
- GPT2-r repeats Joey's excitement many times, while hallucinating a roller coaster that was absent in the original story.
- T5-d: Timex "*After a while*" to beginning of last output sentence ✔ Valid alternative to doing *began* → *had begun* (Choice taken by BART-r)

# Human Evaluation

**Instructions: Plot Preservation Study** (Click to expand)

**Read the instructions given below carefully:**
Read the **two** given stories **S1** and **S2** carefully.
(1) Answer in **Q.1** how well you think **S2 preserves** the **plot/storyline** of **S1**
(2) You can answer on a scale of **1**, which means **least preservation**, to **5**, which means **most preservation**.
(3) By **plot** here, we mean the **actual details** about the **characters**, **situations** and **interactions** between them, the **events** taking place etc which you understand on **reading the full story**.
(4) By **plot**, we don't refer to the exact way the story is written or presented. Its possible for two stories with the **same plot** to be **written differently** - for example, one can be in **formal English** and the other one in **informal English**, or one can be written in **linear** order from **start to end** while the other one **starts from the middle** of the plot and **does a flashback**. Nevertheless, the **actual plot** a **reader understands** from both these written stories would still be **the same**.
(5) We give a **few examples** (Examples 1-3) to help you get a better feel of what we mean in (3) and (4). Read the examples carefully!
(6) Once you are done answering, we ask you to in **Q.2** to write a **short reason** for your decision. Note that this **need not be very long** at all - we just need to get a rough idea of why you thought the plot was preserved, or not preserved.

(a)

**Story S1**

Eddie's dad took him to a wrestling show. Eddie loved the whole show! Afterwards, he got a wrestler's autograph. He decided wrestlers were the best guys ever. Eddie had the best day with his dad!

**Story S2**

Afterwards, he got a wrestler's autograph. He decided wrestlers were the best guys ever. Eddie had the best day with his dad! Eddie loved the whole show! Eddie's dad took him to a wrestling show.

# *NAREOR*: Models

- Base Architectures → GPT2, BART, T5
- Can't simply do prompt-continuation finetuning → Need a way to control for $\mathbf{\Pi}_i$,
- Two simple ways to do this:
  - Encode $\mathbf{\Pi}_i$, using special lettered tags
  - Naively reorder sentences as per  → Pass to the model for actually rewriting in plot-preserving fashion
- Supervised data is scarce - Stage 1 Unsupervised training and Stage 2 Supervised training for all models

# Human Evaluation

- Two studies to together evaluate outputs on 4 aspects
- **Study 1**: Asks annotators to rate stories for fluency, coherence and logical plausibility on a 1-5 scale
- **Study 2**: Compare rewritten target narratives ↔ original story narratives → Is the plot preserved?

# Application II :Challenge Set for Temporal Tasks?

- **Sentence Ordering Task**: Given unordered set of sentences, predict their original discourse order.
- **Intuition**: Non-linear narratives underrepresented in training sets → Test set with all non-linear narratives = distributional shift.
- **Control**: Original stories **S** vs **Challenge**: Human-rewritten target narratives **S'**
- Two models
  - $\mathbf{M_{ext}}$ : Trained on large external SIS dataset
  - $\mathbf{M_{iid}}$ : Trained on ROCStories = original stories S from the $\mathcal{NAREORC}$ train split

# Application I : Interestingness

- How interesting are stories rewritten as per target narrative orders compared to original stories? We ask annotators to answer on 1-5 → 3 = equivalent

- Note that we're simply evaluating with randomly chosen non-linear $\Pi_i$, here → Carefully chosen $\Pi_i$, could potentially be even more interesting!

- Both human and model rewritten stories are evaluated.