Conveying the Predicted Future to Users: A Case Study of Story Plot Prediction

Chieh-Yang Huang,[1] Saniya Naphade,[2*] Kavya Laalasa Karanam,[3*], and Ting-Hao Kenneth Huang[1]
[1] The Pennsylvania State University, PA, USA. {chiehyang, txh710}@psu.edu [2] GumGum Inc., CA, USA. [3] Intel Corporation, CA, USA. * Contributed equally.

Crowd-AI Lab crowdailab.net

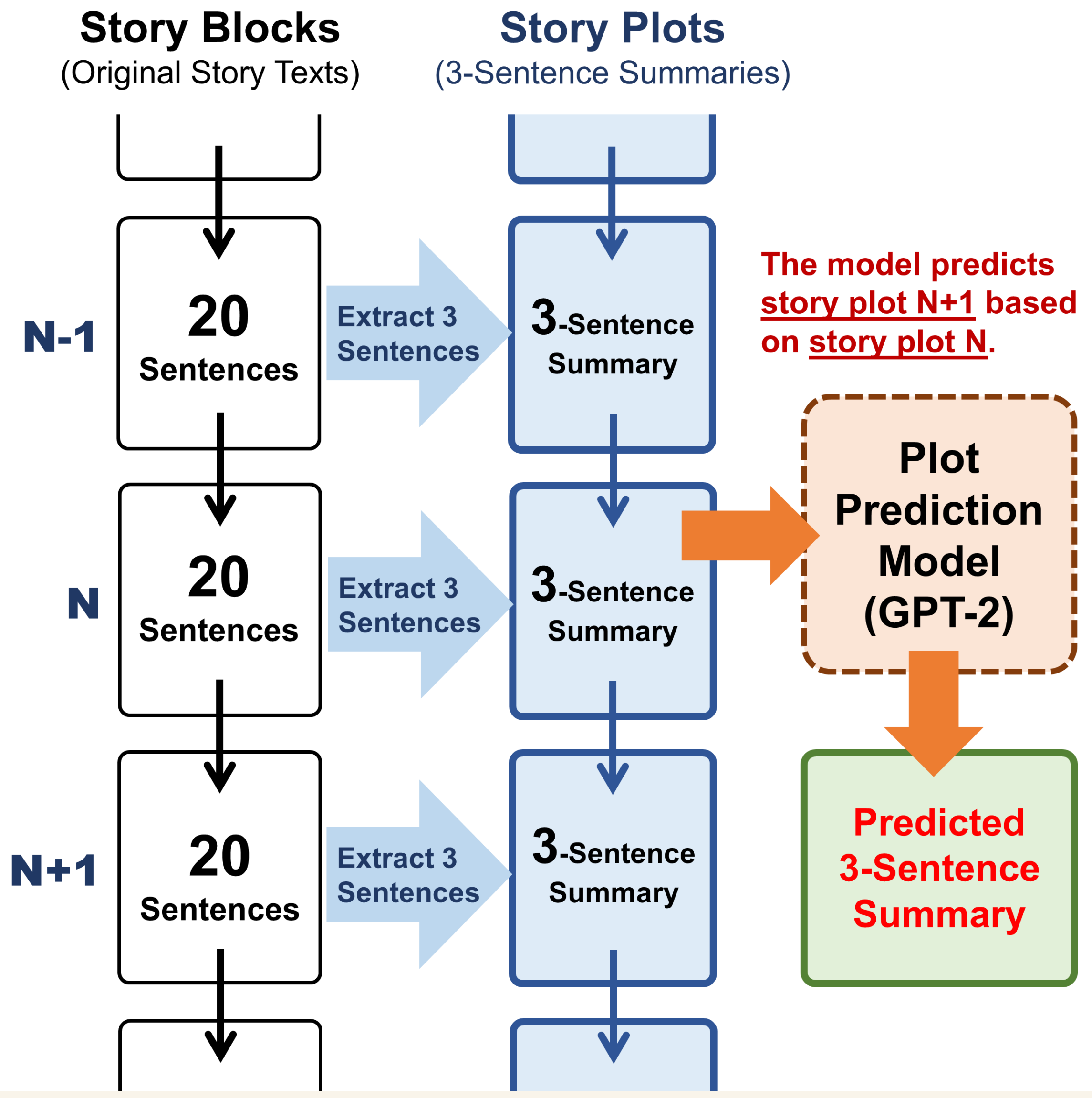
AAAI 2023 Workshop

Introduction

- Story writing is hard. Writers can struggle to develop the follow-up scenes any time.
- Many existing tools are mostly for short stories and **not suitable** for developing long stories.
- LLMs generate content for you directly.

How to support creative writing in practice?

- A long novel = A sequence of fixed-sized (e.g., 20 sentences) story blocks.
- A story plot = A **summary** over a story block.
- **Generate a story plot** for the next story block (i.e., B_{n+1}) using the previous story block (i.e., B_n).



Story Blocks (Original Story Texts)

Story Plots (3-Sentence Summaries)

N-1 20 Sentences → Extract 3 Sentences → 3-Sentence Summary

N 20 Sentences → Extract 3 Sentences → 3-Sentence Summary

N+1 20 Sentences → Extract 3 Sentences → 3-Sentence Summary

The model predicts story plot N+1 based on story plot N.

Plot Prediction Model (GPT-2)

Predicted 3-Sentence Summary

Story Plot Prediction

- Collected story plots by extractive summarization.
  - Used **MatchSum** to collect three-sentence summary.
  - Applied on **Bookcorpus** dataset (900k story blocks).
- Applied three story plot generation models.
  - Fusion-Based Seq2seq [1]
  - Plan-and-Write (P&W) [2]
  - Frame-Enhanced GPT-2 (FGPT-2) [3, 4]
- Other baselines
  - Ground-Truth (GT)
  - Random-History (RH)
  - Random-Future (RF)
  - GPT-3 [5]

[1] Fan, Angela, Mike Lewis, and Yann Dauphin. "Strategies for structuring story generation." ACL 2019.
[2] Yao, Lili, et al. "Plan-and-write: Towards better automatic storytelling." AAAI 2019.
[3] Huang, Chieh-Yang, and Ting-Hao Kenneth Huang. "Semantic frame forecast." NAACL 2021.
[4] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 2019.
[5] Brown, Tom, et al. "Language models are few-shot learners." NeurIPS 2020.

Human Evaluation – Ranking Study for Quality Assessment

Asked workers to rank 2 aspects [6].

- **Consistency**: whether the given story plot **makes sense** in its context (story snippet).
- **Storiability**: whether readers would be **curious** to read the complete story developed from the given story plot.

=> 200 instances * 5 assignments on Mturk

Best GT ≪ FGPT-2 ≪ RF < RH ≪ Fusion-Seq < P&W Worst

Consistency↓	GT	RH	RF	Fusion-Seq	P&W	FGPT-2
Mean Rank	3.091	3.586	3.528	3.733	3.741	3.321
P-values for T-test						
GT	-	<0.001	<0.001	<0.001	<0.001	0.003**
RH	-	-	0.437	0.054	0.039*	<0.001
RF	-	-	-	0.006**	0.004**	0.005*
Fusion-Seq	-	-	-	-	0.915	<0.001
P&W	-	-	-	-	-	<0.001

Best GT ≪ RH < RF < FGPT-2 ≪ P&W < Fusion-Seq Worst

Storiability↓	GT	RH	RF	Fusion-Seq	P&W	FGPT-2
Mean Rank	3.178	3.402	3.452	3.756	3.748	3.464
P-values for T-test						
GT	-	0.003**	<0.001	<0.001	<0.001	<0.001
RH	-	-	0.518	<0.001	<0.001	0.414
RF	-	-	-	<0.001	<0.001	0.877
Fusion-Seq	-	-	-	-	0.915	<0.001
P&W	-	-	-	-	-	<0.001

[6] Roemmele, Melissa. "Inspiration through observation: Demonstrating the influence of automatically generated text on creative writing." ICCC 2021.

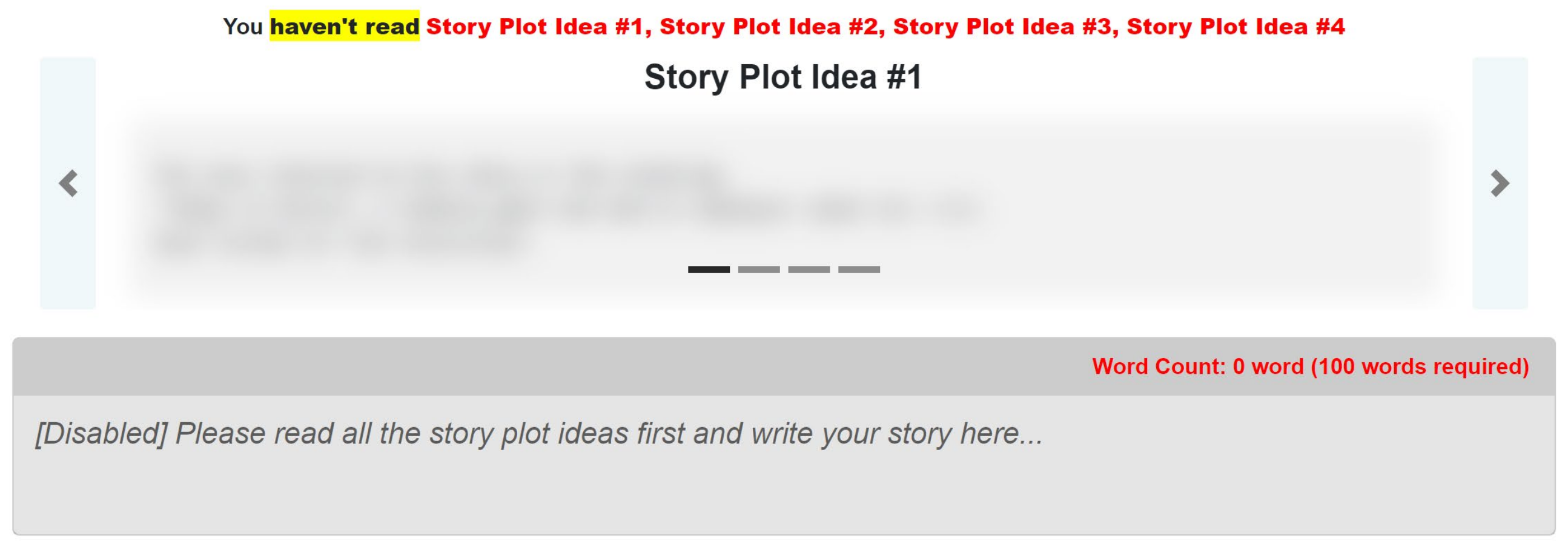
Aspect		GT	RF	FGPT-2	GPT-3
Inspiringness ↑		0.294	0.294	0.176	**0.647**
Helpfulness	Most ↑	0.235	0.353	0.059	0.353
	Least ↓	0.000	0.294	0.294	0.412
	Overall ↑	**0.235**	0.059	–0.235	–0.059
Readability	Easiest ↑	0.353	0.235	0.176	0.235
	Hardest ↓	0.294	0.059	0.471	0.176
	Overall ↑	0.059	**0.176**	–0.294	0.059
Creativity	Most ↑	0.353	0.176	0.000	0.471
	Least ↓	0.176	0.294	0.353	0.176
	Overall ↑	0.176	–0.118	–0.353	**0.294**

Human Evaluation – Writing Study with Story Continuation

Story Continuation Task

1. **Read** story block B_n
2. **Read four story plots** for B_{n+1}
3. **Write a 100-word follow-up story**

=> 5 instances * 5 assignments on Mturk (17 qualified)



You **haven't read** Story Plot Idea #1, Story Plot Idea #2, Story Plot Idea #3, Story Plot Idea #4

Story Plot Idea #1

Word Count: 0 word (100 words required)

[Disabled] Please read all the story plot ideas first and write your story here...

Self-Reported Questionnaire

- GPT-3 is the **most/least helpful** one.
- FGPT-2 is **not effective** in many aspects when compared to other strong baselines.

	GT	RF	FGPT-2	GPT-3	Random
Similarity	0.816	0.795	0.795	0.840	0.787

Semantic similarity between plots and drafts

- FGPT-2 could still influence writing (inspiration-through-observation[6]).

Token alignment

- GPT-3 tokens are not used the most frequently even though having high semantic similarity.

	Story Coverage		Plot Coverage	
	Mean	CI	Mean	CI
GT	0.198	[0.163, 0.233]	0.530	[0.473, 0.587]
RF	0.193	[0.164, 0.222]	0.536	[0.475, 0.598]
FGPT-2	0.163	[0.145, 0.182]	0.484	[0.429, 0.539]
GPT-3	0.170	[0.149, 0.190]	0.498	[0.441, 0.555]
Random	0.151	[0.149, 0.153]	0.450	[0.445, 0.455]