NAACL 2024

Event Causality Is Key to Computational Story Understanding

Yidan Sun[1], Qin Chao*[1,2], and Boyang Li [1]*

{SUNY0053, CHAO0009, boyang.li}@ntu.edu.sg

[1]College of Computing and Data Science,
Nanyang Technological University, Singapore

[2]Alibaba-NTU Joint Research Institute

*Equal Contribution

Paper

Event causality

We say Event A causes Event B if:

- in combination with other factors, Event A is a necessary or a sufficient condition for Event B;
- the occurrence of Event A raises the probability of Event B occurring.

Kicking a ball The ball moving	Winning a lottery Joy
Physical	*Psychological*
The desire for a driver's license Taking the driving test	Misplacing a secret document Losing secrets
Motivational	*Enabling*

Source: Four common categories of event causality [1].

[1] Trabasso, T., Van den Broek, P., & Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse processes*, *12*(1), 1-25.

Event causality offers important information for story understanding

Cognitive science [1,2] indicates that humans rely on event causality in story understanding.

Event causality is utilized in the symbolic approach [3,4] to computational story generation

Mike cheats on his wife

↓ *Enabling*

Mike gets divorced

↓ *Enabling*

Mike has to pay alimony

↓ *Psychological*

Mike is unhappy

[1] Tom Trabasso and Paul Van Den Broek. 1985. Causal thinking and the representation of narrative events. Journal of memory and language, 24(5):612–630
[2] Dennis E Keefe and Mark A McDaniel. 1993. The time course and durability of predictive inferences. Journal of memory and language, 32(4):446–463
[3] Michael Lebowitz. 1985. Story-telling as planning and learning. Poetics, 14(6):483–502.
[4] Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. Journal of Artificial Intelligence Research, 39:217–268

Contribution

Propose a versatile technique for identifying event causal structures:
LLM-prompted Causal Graph Generation

Verify the accuracy of event causal structures on TWO benchmarks:
GLUCOSE and COPEs

Demonstrate the benefits of event causal structures in TWO story understanding tasks:
Story Evaluation and Video-Text Alignment

How to identify causal relations using LLMs[1]?

Prompt:

Here is a list of nodes (events) from a story event graph. We want you to fill in the edges of the event graph with causal connections between nodes. An event graph contains nodes and edges. Each node represents an event, and each edge represents the causal connection between two events.

<Instruction>

Example Input:
Node 0: When Dan goes to school in the morning, he has to take the bus.
Node 1: One day Dan was running late, and missed the bus to school.
Node 2: Dan called his friend Pete, and asked for a ride to school.
Node 3: Pete gave Dan a ride to school, but Dan was late for his first class.
Node 4: Luckily Dan wasn't late for any of his other classes that day.
Example Output:
Edge 0: (Node 0 -> Node 1)
Edge 1: (Node 1 -> Node 2)
Edge 2: (Node 2 -> Node 3)
Edge 3: (Node 1 -> Node 3)
Edge 4: (Node 3 -> Node 4)

< Demos>

Now, it is your turn to construct the event graph for the following event list.
Event List:
Node 0: <S1>
Node 1: <S2>
Node 2: <S3>
Node 3: <S4>
Node 4: <S5>
Output:

<Question>

Edge: (Node A -> Node B)
Nodes: Events
Edges: Event causal relations

[1] ChatGPT-3.5-turbo,Llama2-13B-chat, Falcon-instruction-40B, and Yi-34B-chat.

How to assess its quality?

Benchmarks: GLUCOSE[1] and COPEs[2]

Both are constructed on ROCStories (5-sentence short story)

Sentence 1: The man laid down for a nap.
Sentence 2: His cat jumped on his stomach
Sentence 3: That woke the man up
Sentence 4: The man petted the cat
Sentence 5: The cat took a nap with the man.

Figure 1. Example story from ROCStories

Classification COPEs: identify all sentences that cause the last sentence.

Generation GLUCOSE: list out all causal connections between sentences

Results

	Acc.	Micro F1	Macro F1
	Supervised		
ClozePromptScore	62.06	45.57	58.22
ROCK	66.47	51.90	63.08
COLA	70.29	57.38	67.29
	Few-shot (Ours)		
Falcon-40B-instuct	65.74	41.60	58.68
Llama-2-13B-chat	71.47	47.58	63.99
Yi-34B-chat	72.94	55.98	68.22
ChatGPT-3.5	**74.26**	**57.42**	**69.49**

Table 1: Performance on COPES.

	F1	BLEU	BERTScore	BERT Similarity.
		Supervised		
GPT-2_{large}	59.54	28.92	79.86	84.64
T5_{large}	**61.50**	**31.75**	**84.34**	**88.77**
		Few-Shot (Ours)		
Falcon	28.57	13.43	38.65	25.68
Llama-2	51.70	19.77	58.22	54.82
Yi	57.95	18.95	**77.42**	**84.32**
ChatGPT	**60.75**	**21.20**	75.33	80.89

Table 2: The BLEU, BERTScore, BERT Similarity, and F1 score on GLUCOSE dataset, averaged over dimensions 1 & 6.

Downstream Task 1: Computational Open-end Story Evaluation

Rate the quality of the machine-generated story

Benchmark: OpenMEVA dataset[1]

- Contains 1000 model generated stories in two story domains:
  ROC domain (5-sentence) and WP domain (long stories, 20-sentence)
- Human evaluators rate the quality of the story on a scale from 1 to 5.

Task: Build a scoring system that correlates with human ratings.

[1] Guan, Jian, et al. "OpenMEVA: A benchmark for evaluating open-ended story generation metrics." (2021).

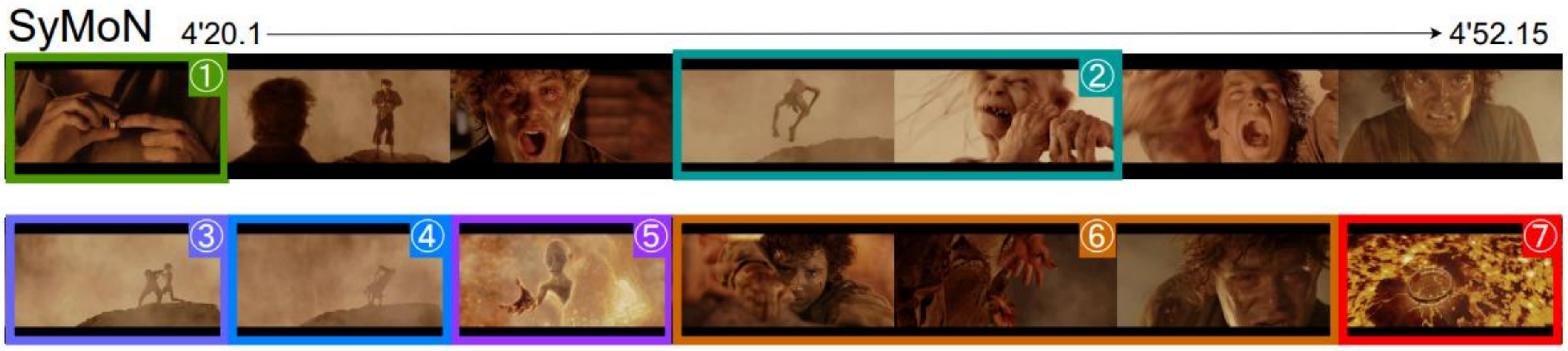
Results

Metrics	OpenMEVA-ROC (n=1000)					
	Writing Prompt Level			Dataset Level		
	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
UNION in-domain*	-	-	-	0.412	-	-
ChatGPT in-domain few-shot						
Repl. Wang et al. (2023b)♠	0.553	0.526	0.466	0.498	0.496	0.398
1 ChatGPT-"causal"	0.560	0.537	0.480	0.501	0.503	0.402
2 ChatGPT-causal-graph	**0.592**	**0.575**	**0.520**	**0.526**	**0.514**	**0.425**

[1]merely insert the word "causal" into the original prompt Corr. increase up to **3%**↑

[2]add the generated causal graph into the updated prompt **3.2% -11.5%**↑

Downstream Task 2: Story Video-Text Alignment

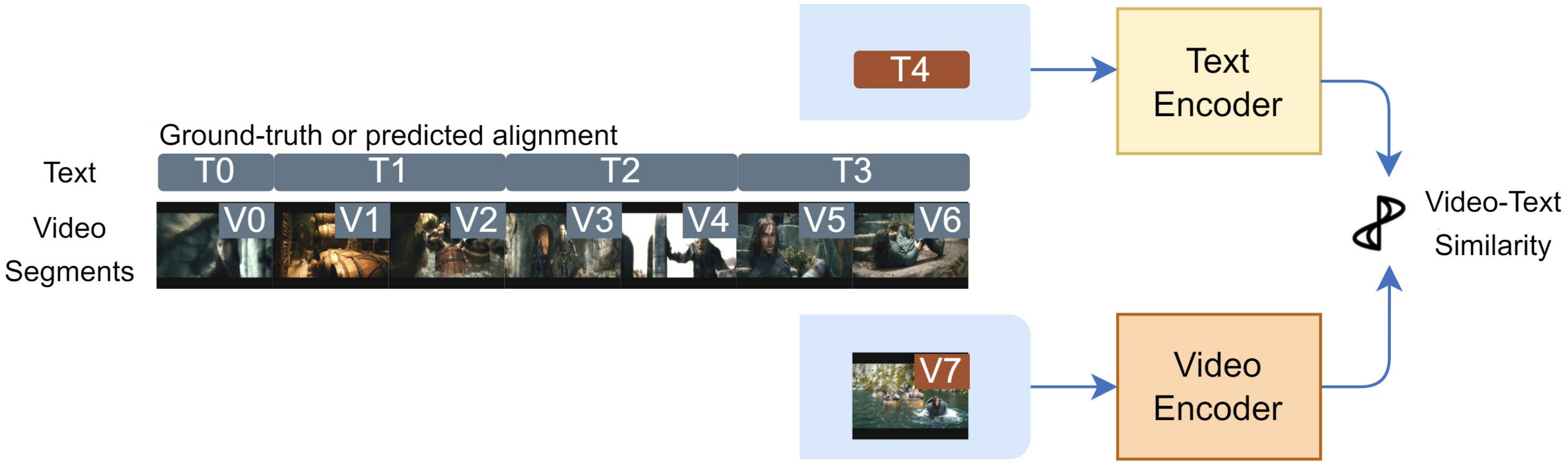


① he succumbs to the ring and claims it as his own, putting it on his finger.② gollum finds the invisible frodo and attacks him, biting his finger off to reclaim the ring.③ frodo attacks gollum in an attempt to reclaim the ring, and in the ensuing struggle, ④ they both fall off the ledge.⑤ gollum falls into the lava with the ring and dies. ⑥ frodo clings to the side of the ledge and is rescued by sam.⑦ as the ring disintegrates into the lava.

Task: Find the best alignment between a sequence of video clips and a sequence of sentences.

[1] Yidan Sun, Qin Chao, Yangfeng Ji, and Boyang Li. 2022. Synopses of movie narratives: a video-language dataset for story understanding

Identify causal context for video-text similarity calculation



Results

	Clip Acc.	Sent. IoU
NeuMATCH Split (sub-sentence level)		
NeuMATCH-MD (Supervised)	4.0	2.4
NeuMATCH-DTW (Supervised)	10.3	7.5
SyMoN-MD	5.9	2.7
Temporal Context-DTW	12.3	7.1
Causal+Temporal Context-DTW	**23.2** (↑**10.9**)	**18.4** (↑ 10.9)
SyMoN Split (sub-sentence level)		
SyMoN-MD	10.1	1.9
Temporal Context-DTW	10.2	8.0
Causal+Temporal Context-DTW	**24.2** (↑ 8.2)	**21.5** (↑**13.5**)

	Clip Acc.	Sent. IoU
NeuMATCH Split (sentence level)		
SyMoN-MD	7.4	3.4
Temporal Context-DTW	29.2	18.3
Causal+Temporal Context-DTW	**33.3** (↑ 4.1)	**22.5** (↑ 4.2)
SyMoN Split (sentence level)		
SyMoN-MD	7.7	3.3
Temporal Context-DTW	32.5	19.6
Causal+Temporal Context-DTW	**40.2** (↑ 7.7)	**27.6** (↑ 8.0)

Locate the antecedents using causal graph and put into the encoder ↑**4.1% -13.5%**

Conclusion

Extract event causality

Solution: LLM + prompt

Assess the quality of LLM-extracted event causality

Verify event causality benefits story understanding

Two Benchmarks in story domain:
GLUCOSE, COPEs

Results:
Set a new SOTA on COPEs

Two Tasks:

1. Story evaluation (subjective)

Result: Correlation ↑ up to 11.5%

2. Text-Video Alignment (objective)

Result:
Acc. ↑ up to 10.9%, Sentence IoU ↑ up to 13.5%

Q&A

Paper

GitHub

Yidan Sun, Qin Chao, and Boyang Li

{SUNY0053, CHAO0009, boyang.li}@ntu.edu.sg