Fine-to-Coarse Entailment Hierarchy Construction for Coarse-to-Fine Story Generation

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Tagyoung Chung
Amazon AGI Foundations

amazon | science

Motivation

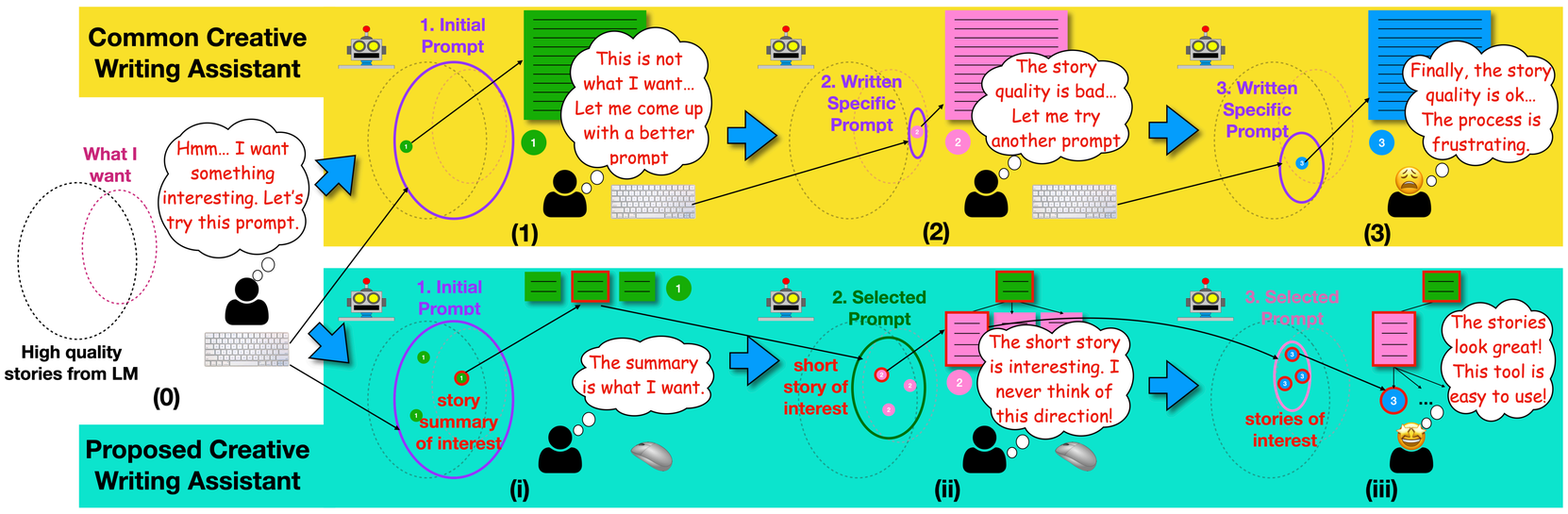

Figure 1: The comparison of existing assistants and our assistant. Each circle refers to a set of stories that satisfy the constraint. (0) a user provides a general initial prompt. Existing workflow: (1) The user reads the story and rewrites the prompt. (2) The LM cannot generate a good story for a new specific prompt. (3) The user feels frustrated after reading two long stories and writing two specific prompts. Proposed workflow: (i) The user chooses one of the generated specific summaries he/she likes. (ii) The system continues partitioning the story space of the user's interest using three short stories. (iii) Finally, the user loves the three stories after reading some short texts and a few clicks.

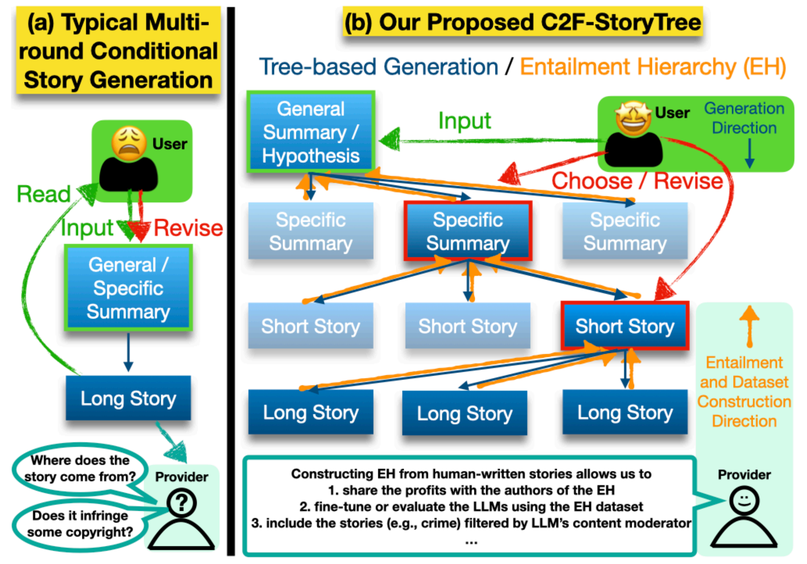

Figure 2: (a) In the typical framework, users often have to repeat the cycle of reading the undesired stories and modifying the prompts in a dialog and the service providers often cannot reduce the risks of infringing copyrights. (b) In our C2F-StoryTree framework, a user can write a general summary/hypothesis, and then our LM provides a more specific summary. The user can choose the desired specific summary as the prompt for the short story, and so on to iteratively expand the story. To realize this framework and give providers more control, we build an entailment hierarchy (EH) dataset and train an LM at each layer to iteratively expand the story prompt.

Method

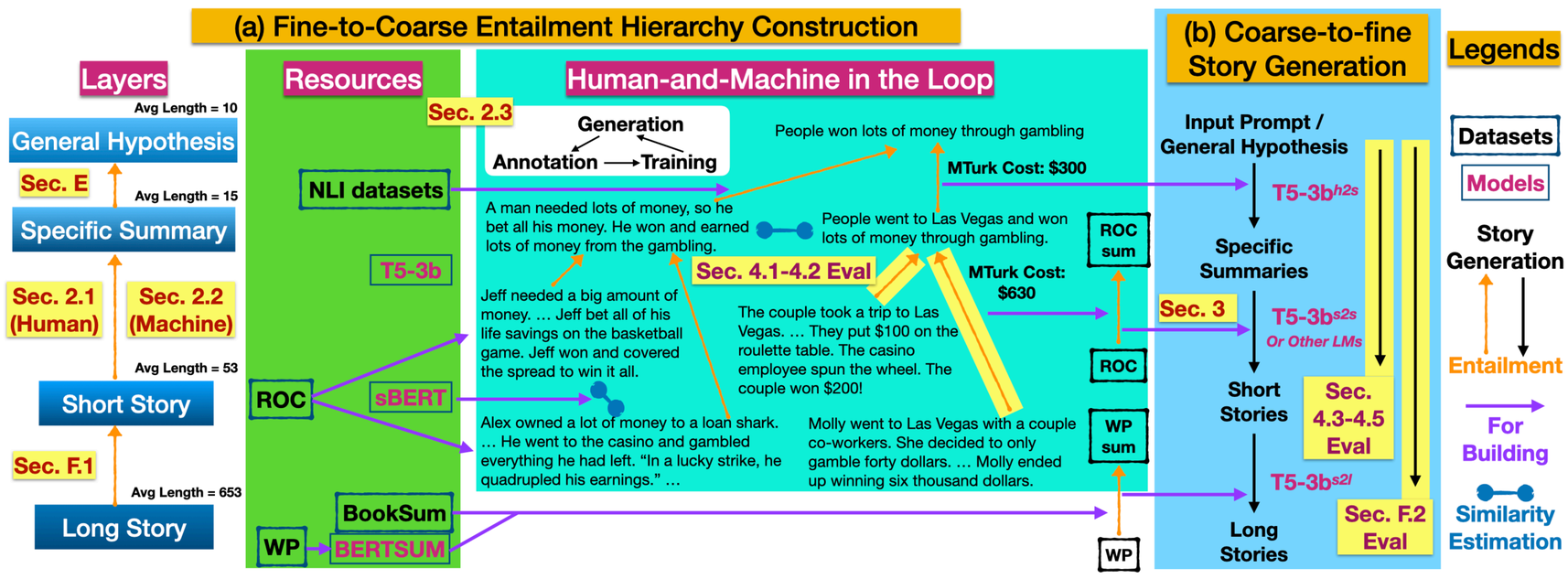

Figure 3: The realization of C2F-StoryTree in this work. (a) We construct the entailment hierarchy (EH) by leveraging existing resources and machine-and-human-in-the-loop techniques. After fine-tuning the models using the crowdsourced EH, we generate summaries for ROC and WritingPrompts (WP). (b) We train seq2seq models to generate lower-layer text from upper-layer text and achieve the coarse-to-fine (C2F) generation.

General Summary/Hypothesis as the Input Prompt: "A girl achieved something impressive."

Short Story from CG (Conditional Generation): "A girl was very tall with a long teeter totter. But she wanted to do something with it. Her parents bought her a teeter totter. The girl sat on the stool and stared at it intently. She had finally completed something that would be impressive to her mom."

Stories from EH + rerank (T5-3b FT)

Specific Summary from T5-3b^{h2s}		**Short Story** from T5-3b^{s2s}		**Long Story** from T5-3b^{s2l}
Text	RS	Text	RS	Text
1. A young girl did something extraordinary during the school year.	0.79	1. Amy had a test on Friday. She was very scared. She decided to jump rope. Amy threw the rope very hard. Amy got a straight A.	0.99	1. Amy was a nerd, a nerdy teen. ... So when she heard about the test she was scared. She figured, she would have to jump rope. ... She threw the rope and she got a straight A! She was so happy!
2. A girl had a great achievement on her birthday.	0.77			
3. A girl has become a runner after running a half marathon.	0.73			
4. A girl did something impressive at the end of the day.	0.65	2. A gymnast went to her gymnastics team practice. ... She got on the trampoline and started to flip. ... Her coach was proud.	0.98	
5. A student made an impressive effort to reach her goal.	0.25			
6. Someone had an experience with achieving something at school.	0.03			

Table 1: An example generated by our C2F-StoryTree framework and the conditional generation baseline. We highlight the selected text for generating more specific text. RS refers to the reranker scores.

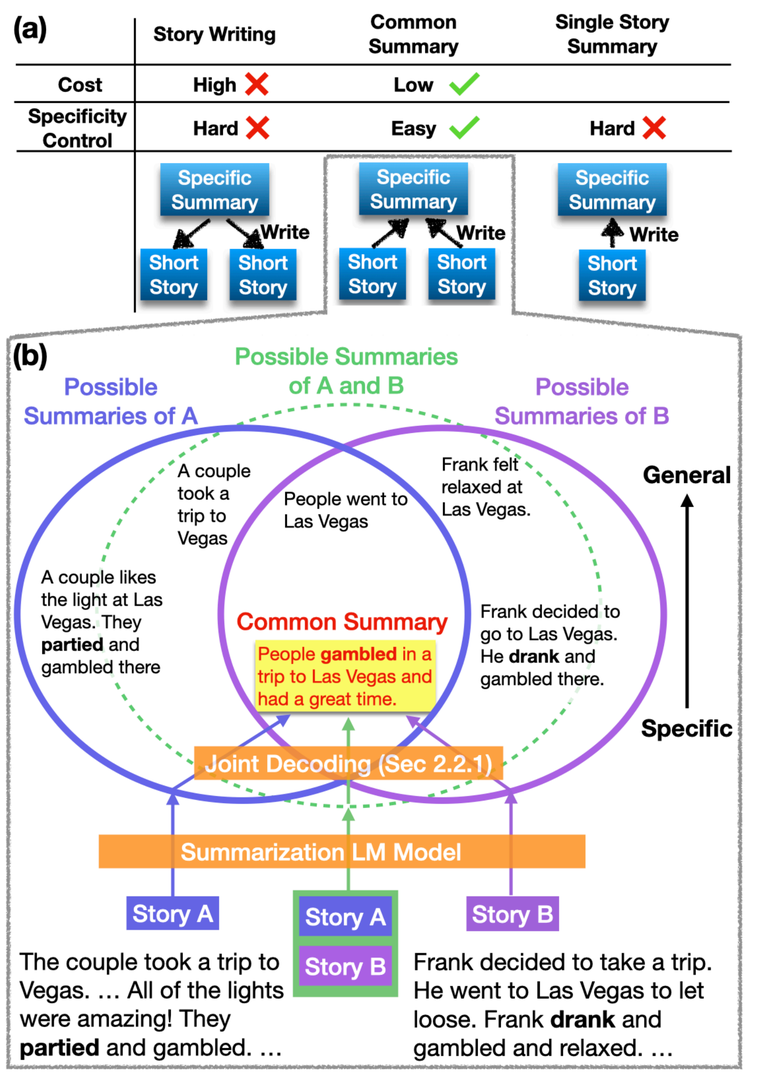

Figure 4: (a) Comparison of options for constructing the entailment hierarchy. (b) The task and our model of generating a common summary. We input story A, story B, and their concatenation into a summarization model and during the decoding time, we merge their generation probabilities word by word in order to get the common summary of both story A and story B.

Experiments

Method	Story Generator	C2F	len	Reference Relevancy (%) R1	R2	sim	Creativity (%) R1 (↓)	R2 (↓)	sim (↓)	Coherence ppl (↓)	Diversity (%) dist-1	dist-2	Human Judgement Rel	Pro (%)	Cr	Coh	Eng
Round for Generators with Open-Source Licenses																	
CG	T5-3b FT		46.61	25.20	4.11	45.49	24.48	11.92	53.56	10.00	37.73	73.03	2.44	83.33	2.82	3.22	2.72
EH	T5-3b FT	V	47.94	23.70	3.03	43.51	**14.78**	**2.90**	**44.12**	**9.97**	**42.61**	**81.63**	2.48	71.67	**3.11**	3.81	2.88
EH + rerank	T5-3b FT	V	47.70	25.07	3.77	46.64	16.79	3.98	48.61	**9.97**	40.63	78.66	2.68	**92.50**	3.10	**3.88**	**2.97**
EH + rerank + sim	T5-3b FT	V	47.23	**25.71**	**4.25**	**49.27**	17.62	4.41	48.96	9.99	39.28	76.92	**2.76**	91.67	3.08	3.87	2.78
EH	GPT-J 6B FT	V	64.00	22.90	2.77	40.24	12.50	2.28	38.07	9.34	41.73	81.89	2.37	46.67	3.51	3.83	3.50
Round for Generators using Proprietary Data																	
CG	Vicuna 7B 5 Shot		46.21	23.77	3.21	43.79	20.68	6.98	50.74	10.01	**42.54**	80.47	2.55	**88.33**	3.10	3.99	3.06
EH + rerank	Vicuna 7B 5 Shot	V	49.85	**24.67**	3.56	45.65	17.74	4.16	50.59	9.86	41.43	80.83	2.59	86.67	3.28	**4.16**	**3.38**
EH + rerank	Vicuna 7B FT	V	52.33	24.43	**3.68**	**45.72**	**15.56**	**3.32**	**46.39**	**9.79**	42.37	**81.54**	**2.72**	80.00	**3.56**	3.98	3.36
CG	GPT3 175B 5 Shot		58.37	25.15	3.46	47.57	18.43	5.50	54.82	9.55	38.93	76.97	2.95	90.00	3.42	4.13	3.59
CG	GPT3.5 175B 5 Shot		57.91	24.94	3.46	47.26	18.49	5.37	54.41	9.57	39.71	78.26	3.02	90.00	3.67	4.35	3.63
ROC NN Stories	Human		50.87	27.75	4.50	61.56	11.87	1.42	36.95	9.86	46.38	86.47	2.78	54.17	3.52	4.20	3.58

Table 2: Comparison of generated short stories. Our coarse-to-fine (C2F) generation framework selects the generated option that has a high reranker score (EH + rerank) and our main baseline is a conditional generator without using entailment hierarchy (CG). FT means fine-tuning, Rn is ROUGE n F1, and sim is the similarity measured by sBERT. We report human judgments on the reference relevancy (Rel), prompt following/entailment probability (Pro), creativity (Cr), coherency (Coh), and engagement (Eng), which are not directly comparable across the two rounds. The best scores using a T5-3b or Vicuna 7B model are highlighted. GPT3 refers to *davinci* and GPT3.5 refers to *text-davinci-003*.

Method	len	Automatic Metrics Fluency ppl (↓)	Diversity (%) dist-1	dist-2	Human Judgement Entail-2 (%)	Entail-1 (%)	Fluency	Specificity (↓) (%)
GPT3.5 (text-davinci-003) 1 Shot	19.27	9.51	36.86	64.90	44	66.3	4.58	**23.5**
GPT3.5 (text-davinci-003) 5 Shot	15.87	**9.05**	37.98	**69.17**	45.5	66.5	**4.69**	25.5
DynE (Hokamp et al., 2020)	11.56	10.08	**41.41**	67.04	25	35	4.45	26.5
JD	12.16	10.01	41.20	66.85	25	37	4.53	24
JDC	11.05	10.19	41.04	65.32	33	47	4.53	26.5
JDC + rerank	10.64	10.24	33.18	54.17	58	69.5	4.66	38
JDC + rerank + hypo	10.04	10.30	34.29	54.89	**72**	**80.5**	4.68	48

Table 3: Comparison of common summary generation methods. We use T5-3b models trained by the entailment hierarchy in dynamic ensemble (DynE) (Hokamp et al., 2020), joint decoding without concatenation (JD), joint decoding with concatenation using Equation 1 (JDC), our method with reranker (JDC + rerank), and our method trained by all summary/hypothesis layers (JDC + rerank + hypo). Entail n means the probability that the summary is implied by n stories. len is story length. ↓ means lower is better and the best scores are highlighted.

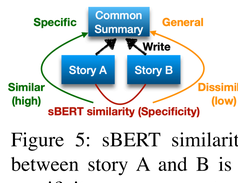

Figure 5: sBERT similarity between story A and B is a specificity measurement of their common summary.

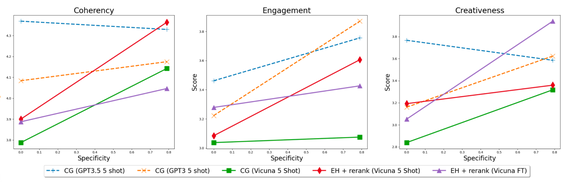

Figure 6: The linear trendlines of the short story scores from different LLMs given the prompt with different specificity levels.

Conclusions

- We propose C2F-StoryTree, a coarse-to-fine tree-based story generation framework. Compared to the typical story generation workflow, our framework is more controllable and transparent and can lead to high-quality stories that align well with the user's interest.
- We propose a cost-efficient bottom-up way to build an 'entailment hierarchy', a text data structure for realizing the framework, using human-and-machine in the loop and existing public datasets in the creative writing domain.
- We propose a new common summary writing task, which avoids the resulting summary/ prompt from being too specific or general. In the task, we can also measure the specificity of the prompt using the similarity of the two input stories, which allows us to test the performance of different models given different prompt specificities.
- We propose a novel method to generate a common summary of two stories using joint decoding and reranker. We demonstrate that our finetuned model is much better than GPT3.5 few-shot in-context learning.
- We design a series of MTurk tasks to mini- mize the cost and maintain the high annotation quality. Our method allows us to build a high- quality 2-layer entailment hierarchy on top of 5-sentence stories by spending less than $1000 on MTurk.
- We design and conduct comprehensive evaluations to compare the quality of common summaries and stories from different generation systems in various aspects.