SSP: Story-Space Prompting Improves the Reader Immersion in Long Story Generation

Anonymous ACL submission

Abstract

Generating long-form stories with neural network models, even the large language models (LLMs), e.g., GPT, has always been criticized for lacking interestingness and coherence, thus greatly diminishing the reader's sense of immersion. In this paper, we present a novel "story space" prompting (SSP) solution, which provides a coherent and consistent background to support long-term storytelling. Specifically, we first define the story space intricately connected to the given story premise. Then, Our framework systematically generates the story space by progressively constructing it from an abstract representation to a more informative and detailed one. Empirically, we implement our plug-in method upon an existing advanced story generation framework (Yang et al., 2023) and evaluate its impact on both interestingness and coherence. Our findings emphasize the significance of our SSP in enhancing reader engagement and immersion, contributing to advancements in long-form story generation.

1 Introduction

Advancements in natural language generation systems have sparked a growing interest in long-form text generation, where texts extend over thousands of words or more (Cho et al., 2019; Tan et al., 2021; Guo et al., 2022). Crafting long stories is a formidable task as it must ensure a coherent and consistent plot that extends over thousands of words, staying true to the initial premise. Additionally, it must maintain the narrative style throughout the entire story and prevent any factual contradictions, which become more challenging to manage over an extensive narrative horizon.

Recent efforts, notably the Re^3 method (Yang et al., 2022), have pioneered the generation of extended narratives by iteratively refining and expanding stories through a recursive process. Subsequently, DOC (Yang et al., 2023) enhances the outlining phase within the Re^3 framework. These approaches offer a potential solution to improve complex long-form content. Despite advances, current state-of-the-art story generation methods, e.g., DOC, still underperform human capabilities in several aspects, notably in creating ***immersion***: the AI-generated stories often depict protagonists taking actions in a seemingly empty world, lacking the immersive elements that would fully engage the reader within the story world. An immersive story is always accompanied by a rich background (Gander, 1999), which is crucial for story development, influencing its tone and overall effective (Rozelle, 2005). Normally, rich backgrounds are built based on interesting and multi-dimensional story spaces. For example, J.R.R. Tolkien particularly constructed the captivating world of Middle-earth in The Lord of the Rings.

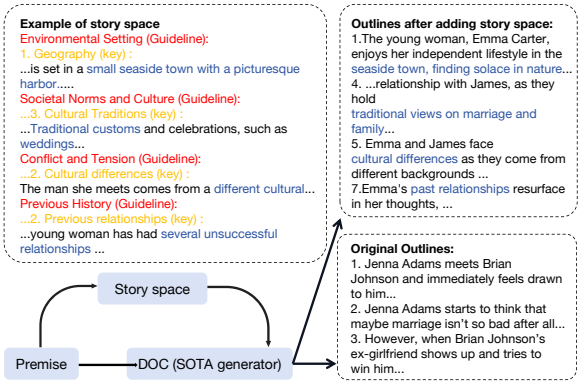

Figure 1: A case study of concatenating **SSP** with **DOC** system. The red words are the guidelines, The yellow words are the keys, and the black words are the value, respectively, in our story space

Motivated by above observations, we propose to improve the reader immersion of the AI-generated long story by introducing a plug-and-play "story space prompting (SSP)" mechanism upon the SOTA story generation framework (Yang et al., 2023). The philosophy of our proposed mechanism is to enhance the structure and continuity of the

story background conditioned on a given premise. To achieve this, we ❶ manually design the guidelines that define abstract dimensions spanning the story spaces (e.g., "*Environmental Setting*" in Figure 1), making them applicable to various story genres; ❷ automatically generate keywords that are highly correlated with the given premise (e.g., "*Geography*" in Figure 1); ❸ carefully develop an agent capable of real-time retrieval from extensive database to supplement the corresponding knowledge value for each keyword, and reinforcing the connection between keywords and premise (e.g., "*a small seaside town with a picturesque harbor and sandy beaches...*" in Figure 1).

To assess the effectiveness of our strategy, we integrate our SSP with the DOC during the generation of the story outline. By combining our SSP with DOC, we evaluate the impact of the story space on the overall quality and coherence of the generated narratives. Experimental results show that SSP achieves substantially higher plot coherence (17.9% absolute increase), interestingness (22.1%), and perceived immersion (22.8%) when judged by human annotators in pairwise evaluations (see Section 4).

In short, our main **contributions** are as follows:

- We define the story space serving as the story background in the task of long-text story generation, and we proposed our SSP framework to automatically generate story space according to the given premise.
- We implement our framework by concatenating with the previous long-text story generator. The experiment results show that our method significantly improves interestingness, coherence, and therefore immersion.

2 Related Work

There are two categories in story generation with external knowledge. The first category is transforming structured knowledge data like knowledge triples into natural language texts. Guan et al. (2020) targeted on generating a commonsense story given a starting sentence. Xu et al. (2020) proposed MEGATRON-CNTRL(Speer et al., 2018), which follows that transforming knowledge triples in Concept-Net into respective sentences by a template and selecting knowledge-conditioned sentences generated by GPT-2(Radford et al., 2019) into the story.

The second category is building a knowledge reasoner to guide the generator. C2PO (Ammanabrolu et al., 2020) utilizes COMET (Bosselut et al., 2019) to generate intermediate events between any two consecutive plot points that infill the plot lines of a story are extracted. CAST (Peng et al., 2022) aims at generating character-centered stories where each sentence should contain one or two characters. It utilizes COMET as an inferential instructor to supervise the story generation process.

3 Methodology

3.1 Setup

In this task, a given or generated premise p is necessary, it's normally a 30-50 words short paragraph. Our ultimate goal is building a story space $\mathcal{S}$ is highly correlated to the premise p. This story space based on p is spanned by a set of independent guidelines g for one direction, like "temporal setting" and "societal norms etc. For each guideline, it's designed as a vector $(k_1 : d_1, k_2 : d_2, \ldots)$ consisting of some key-value pairs, where the key k_i stands for a specific implementation of the premise in this direction, and the value d_i is corresponding detailed descriptions. Figure 2 shows the example of the story space on the mentioned premise.

3.2 Frame work

To make the corpus closer to the premise, we choose to hierarchically generate this corpus:

Story Genres

At the initial stage of the framework, the decision of the genre plays a crucial role as it significantly impacts the development of the story's background. For instance, as depicted in the upper left quadrant of figure 2, the ultimate choice made by the female character is influenced by the chosen story genre. In the case of a "Romance story" genre, she may decide to marry the man, whereas in a "Realistic fiction story" genre, her decision may be influenced by the discovery of certain secrets about the man.

Guidelines

The guidelines devised in this framework are characterized by their high level of abstraction, allowing them to be applicable across various story genres. They are intentionally designed to be independent to avoid repetition of key elements. Each guideline serves as a guiding principle for LLM to generate premise-related keywords that align with its specific direction. The guidelines have

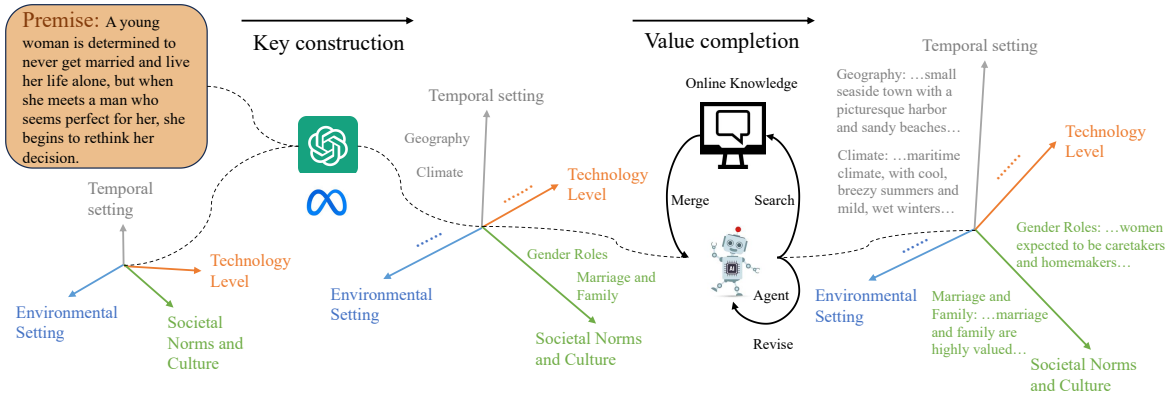

Figure 2: Flow chart of our proposed SSP.

been meticulously crafted and designed by humans. Eventually, there are a total of 7 guidelines to provide meaningful dimensions for the construction of the story space. In figure 2, we only demonstrate 4 directions for spanning the space of the story world.

Key construction

The keys in each guideline should be related to the premise. We generate keywords as follows:

$$(k_1, k_2, \ldots)_{pg} = \text{LLM}(p, \text{prompt}_g)$$

where prompt_g is a prompt designed in a guideline g. To mimic the brainstorming, we sample n times on the key generated distribution, where n is a hyperparameter, and make a selection to make sure the results are desired. The selection methods are as follows: Inspired by the method proposed by Weng et al. (2023) which concludes that LLMs have a strong ability to self-verification, allowing them to assess the conclusions they generate accurately. The first selection method is simply putting all of the generated keys into the LLM and getting the selected keys.

Value completion

After constructing our desired key on a guideline, we need a detailed description corresponding to each key. For instance, in Figure 1, after we generate a key "Geography" in the guideline "Environmental Setting", we need a detailed description, "a small seaside town with a picturesque harbor and sandy beaches", to explain the correlation between the key and the premise. The requirement for this description should also be informative and based on real-world knowledge. We extend the generations in Part 3.2 with value initialization.

$$(k_1 : v_1, k_2 : v_2, \ldots)_{pg} = \text{LLM}(p, \text{prompt}_g)$$

Real-time retrieval

We adopt a similar approach to the paradigm introduced by Yao et al. (2023) known as ReAct, which combines the ability of LLM at reasoning and acting. In our case, we develop an agent specifically designed for value completion. The three available choices are as follows: (1) Search: If we choose the "Search" action, we will conduct a search using an entity generated by LLM (Entity $= \text{LLM}(\text{prompt}, v_i, k_i, p)$, where i stands for i-th key-value pair) to further improve the existing knowledge. Differing from ReAct, which relies on searching for entities on Wikipedia and reasoning based on the search results, our agent takes a different approach. It retrieves knowledge from both the commercial system Bing Chat and Wikipedia. (2) Revise: Alternatively, we can choose the "Revise" instead of the original reasoning part, which involves making revisions to the existing knowledge in order to make it more closely correlated to the premise inspired by the discovery of self-improve ability proposed by Fu et al. (2023). (3) Finish: "Finish" action indicates that LLM believing that the current knowledge is sufficient and that no further modifications are required.

4 Evaluation

The generation of stories and the components within our framework cannot be easily defined by rule-based constraints. Instead of imposing rigid rules on acceptable outputs, we adopt a more flexible approach and evaluate the generated content using human-annotated metrics.

4.1 Agent

Due to online knowledge retrieval, the value generated with the agent has more entity information

and it is revised by LLM to have a closer distance from the premise. As shown in Table 1, annotators think that the value with the agent has more informativeness. We find that LLM may provide some entities that do not exist so the Wikipedia wrapper gets some irrelative information. This may affect the LLM to self-improve so that the final value does not have a satisfactory result.

Method	Informativeness
Initial	17.12
Initial w. Agent	**87.82**

Table 1: Pairwise comparisons of Agent refining against baselines (initial value) from 20 stories. Bold indicates significance with $p < 0.05$. Refined values are rated with more information

4.2 Final story Setup

For the main results comparison, our setup is similar to DOC. The input is an English premise and the output is a corresponding story generated automatically. We adopt 20 diverse premises in DOC as a testset. To decrease noise, we compare 2000-2500 word stories corresponding to the same premise. We use three main metrics, also similar to those from DOC.

- *Interesting* The percentage of passages judged interesting.
- *Coherent* The percentage of passages judged plot-coherent by human annotators.
- *Immersion* The percentage of passages judged as having an immersion by human annotators.

Baselines

Our primary baseline for story generation is based on the DOC system proposed by Yang et al. (2023), which is known to automatically generate stories of similar length. DOC is an extension of the Re^3 system, but it includes an outline controller that allows for longer outlines to be generated. In our analysis, we explore different ways of incorporating our story space into the DOC system.

Results

The results in Table 3 confirm that our framework does substantially improve the ability of DOC to make the story more interesting and under the control of our space, the plot is more coherent. The generated story in Table 2 serves as a concrete illustration of how the outlined guidelines are effectively implemented. The story demonstrates a clear alignment with our story space, indicating a degree of control in the narrative. Although our framework concatenating DOC achieves a better immersion for the reader, the overall numbers of immersion remain low, pointing to two issues. First, the current concatenating method could not utilize all the information in the story space. Second, our story space may be not sufficient or correlated due to the problems encountered.

GENERATED STORY:

Emma Carter had always cherished her independent lifestyle in the serene seaside town of Willowbrook. With the sound of crashing waves as her symphony and the salty breeze as her constant companion, ...[954 words]...

Emma's unwavering commitment to her independence drew concern from her family, particularly her parents. They held staunchly traditional views on marriage and family, ...[104 words]...

One evening, as they strolled along the shoreline, the topic of marriage emerged. Emma expressed her reservations about the institution, sharing her belief that it could...[419 words]...

In the stillness of her seaside home, Emma would sift through memories like grains of sand slipping through her fingers. She recalled the times she had opened her heart, only to have it shattered by betrayal and disappointment. ...[195 words]...

Table 2: An example of generated stories by the given premise

Method	Interesting	Coherence	Immersion
DOC	40.3	45.3	34.7
DOC w. SSP	**62.4**	**63.2**	**57.5**

Table 3: Pairwise comparisons between DOC with SSP and the baseline DOC were conducted for 20 stories. Bold indicates significance with $p < 0.05$. The stories generated using SSP are rated substantially more interesting, coherence and immersive compared to the baseline

4.3 Conclusion

In the task of long-text story generation, the concept of the "story space" serves as the narrative backdrop, providing context and coherence to the story. To address this, we introduce the SSP framework, which automates the generation of the story space based on a given premise. Our approach involves integrating the SSP framework with an existing long-text story generator. Experimental results demonstrate that our method notably enhances the interestingness, coherence, and overall immersion of the generated stories.

Limitations

Indeed, evaluating the quality of story outputs in long-form text generation is a challenging task that often requires human annotations. While human annotations can be expensive and time-consuming, the difficulty of evaluation limits us from running more detailed ablations. Many of our prompts are carefully engineered in English cause two limitations: the process of prompt engineering involves human intervention, which can introduce unintentional bias. Meanwhile, designing carefully engineered prompts can be time-consuming and resource-intensive.

Ethics Statement

We place great importance on ethical considerations and adhere strictly to the ACL Ethics Policy. While we recognize the potential for harm associated with powerful language generation models, our specific focus on generating fiction stories helps mitigate certain ethical concerns compared to broader language models. By emphasizing the generation of fictional content, we aim to minimize the potential for generating toxic or untruthful text that could be used maliciously. we believe that this research will not pose ethical issues.

References

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. In *AAAI*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. 2019. Towards coherent and cohesive long-form text generation. In *WNU*.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint*.

Pierre Gander. 1999. *Two myths about immersion in new storytelling media*. Lund University.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. Knowledge-enhanced pretraining model for commonsense story generation. *TACL*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *NAACL*.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2022. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of EMNLP*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

R. Rozelle. 2005. *Write Great Fiction - Description & Setting*. Write Great Fiction.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint*.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *NAACL*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of EMNLP*.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *EMNLP*.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *ACL*.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *ENNLP*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.