Event-Based and Graph-Based Approaches to Skeleton Selection for Goal-Driven Storytelling

Anonymous ACL submission

Abstract

This study addresses the task of selecting a skeleton for narrative story generation from closely-associated Story Plan Graphs (SPGs). While advanced language models, such as Large Language Models (LLMs), demonstrate potential, they often fall short in manifesting semantic consistency. Utilizing the SPGs generated by Neural Story Planning, which ensures the logical soundness of symbolic planning, we introduce two novel methodologies for skeleton selection: event-based and graph-based approaches. These methods discern salience events within the fabula, helpful the selection of skeletons that are engaging, coherent, and logically consistent. Evaluated against the GPT-3.5 using the ROCStories dataset, our approach evidences enhanced skeleton selection capabilities, offering an efficient and cost-effective solution for skeleton selection.

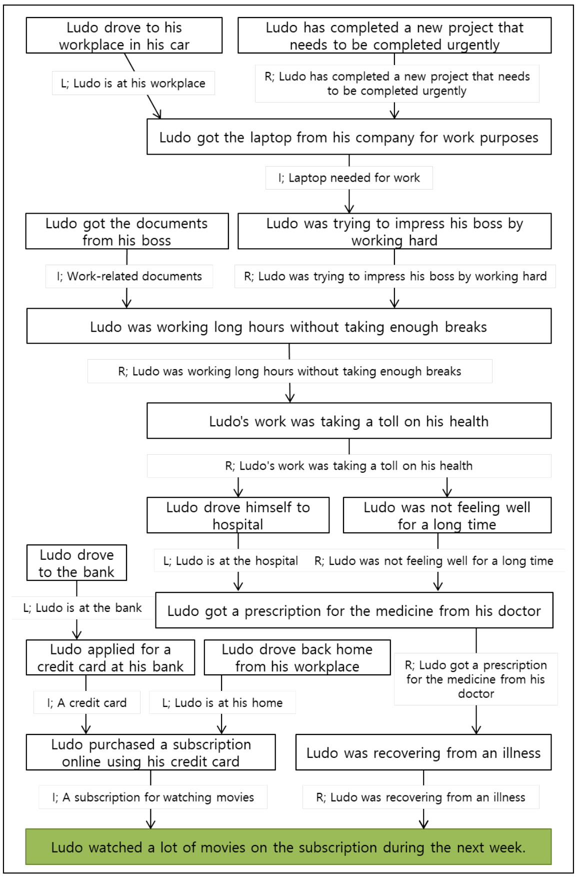

Figure 1: An illustrative example of a SPG generated through the Planning method. It is constructed from connections between event sentences and condition sentences. The green box represents the goal event sentence, the white boxes denote event sentences, and the inscriptions on the connectors signify condition sentences. Within the condition sentences, 'I' stands for item need condition, 'L' represents location condition, and 'R' indicates reason condition.

1 Introduction

Stories are an essential element that permeates human culture and history. They are expressed in various forms, literature, movies and entertainment such as games, providing enjoyment to people. Among them, a narrative story refers to a series of events linked by causality and experienced/generated by actors (Bordwell, 1980; Bal and Van Boheemen, 2009). For instance, a statement indicating a state like "My dog has fleas." is non-narrative, while a statement consisting of an event such as "My dog was bitten by a flea." is considered narrative (Abbott, 2002).

Narrative story generation distinguishes itself from other text generation tasks in that the generated sentences require closely-knit semantic associations, making it notably challenging. Furthermore, crafting a story that engages and entertains the reader presents an additional imperative. Such generated narratives hold potential applications in various entertainment sectors, aiding authors and creators in story ideation and expansion. Consequently, narrative story generation has captured the interest of researchers for decades and continues to be a topic of intensive investigation, especially with the advent of Transformer-based language models (Vaswani et al., 2017; Radford and Wu, 2019) and

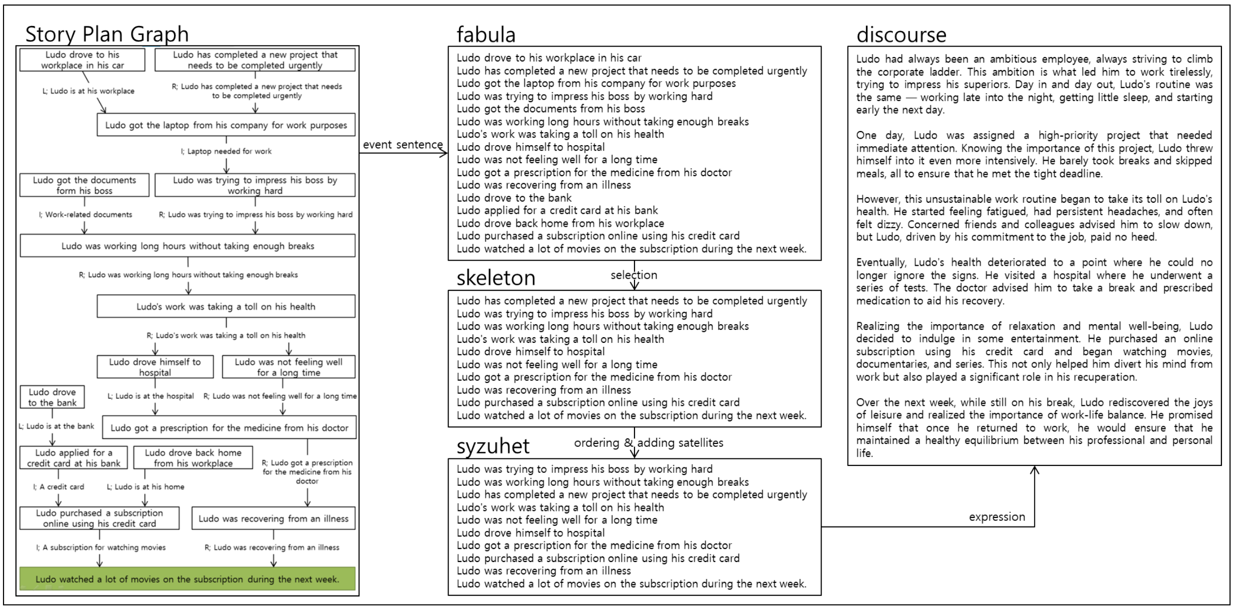

Figure 2: Depicts the progression from a SPG to discourse. For a detailed explanation, refer to the 'Tripartite Models' under Section 2.

the advancements in natural language processing enabled by LLMs. However, there are inherent challenges with these language models in effectively modeling semantic dependencies between sentences. They also manifest limitations in narrative generation, such as repetitions and generic responses (Holtzman et al., 2019).

Traditional solutions to story generation, such as symbolic planning (Meehan, 1977; Lebowitz, 1985; Porteous and Cavazza, 2009; Riedl and Young, 2010; Ware and Young, 2011), can infer causal relationships between events and ensure the logical soundness of stories generated through these relationships. Additionally, it offers a means to model semantic dependencies in the form of a graph. However, symbolic planning requires a hand-crafted schema library that dictates which events can be used. Recent studies have sought to mitigate this limitation by integrating symbolic planning with LLMs in an approach termed Neural Story Planning (Ye et al., 2022). While the modeled SPG (refer to Figure 1) allows for the composition of a story using all event sentences present within the graph, a more consistent and engaging narrative story (or discourse) is formulated by undergoing a process as depicted in Figure 2.

We aim to concentrate our research on the method of selecting a skeleton from the event sentence list (fabula) within the SPG generated by Neural Story Planning. Specifically, this study investigates which events should be chosen from the provided SPG to construct the most effective skeleton leading up to the goal event. The criteria for an optimal selection are defined by the coherency, logicality, and interest of the resulting skeleton. This definition draws inspiration from narrative psychologists who underscore the significance of plot consistency in reader comprehension and story reception(Trabasso and Van Den Broek, 1985; Graesser et al., 1991). To address this challenge, we categorize events in the fabula as either event-based or graph-based to determine their relevance and subsequently perform skeleton selection based on these evaluations. We assessed the effectiveness of this system using the ROCStories dataset (Mostafazadeh et al., 2016) as the basis for the SPGs and fabulas, and compared its performance with the high-performing GPT-3.5 using an A/B testing approach. The results demonstrated that our skeleton selection method outperformed GPT-3.5 in terms of superior selection capability.

The key contributions of this research are enumerated as follows:

- We propose two methodologies leveraging the SPG to select a skeleton that is interesting, logical, and coherent.
- Our skeleton selection approach enhances the intrigue of the selected skeleton through an event-based methodology and ensures its logicality and coherence through a graph-based approach.

- Our selection procedure builds upon well-established significance metrics, namely TF-IDF (Sparck Jones, 1972) and PageRank (Page et al., 1998), obviating the need for substantial computational cost in the selection process.

- We conducted an automated evaluation using GPT-3.5, and the results indicate that our selection technique yields a more optimal skeleton.

2 Background and Related Work

Tripartite Models (Chatman, 1978; Genette, 1983; Rimmon-Kenan, 2002; Walsh, 2001). The '*fabula*' refers to the comprehensive story world, encompassing all events, characters, and circumstances. In this paper, the event sentence list from the SPG was utilized as the fabula. All events within the fabula are feasible, distinguishing it from the '*possible world*' (Ryan, 1991), wherein not all possessed events can occur concurrently. The '*skeleton*' is derived by selecting only the pivotal events from the fabula, essentially constituting the backbone or the primary events ('*nucleus*') of the story. The '*syuzhet*' is responsible for ordering the nucleus of the skeleton to instill elements such as suspense, thereby captivating the audience; it may also incorporate '*satellites*'—events that might not be crucial to the storyline but are pivotal for narration. The '*discourse*' represents the syuzhet as expressed through mediums like text or film. This phase can be regarded as the final narrative story manifested through a specific medium. Our research was primarily concentrated on skeleton selection, grounded in the aforementioned theories and definitions.

Neural Story Planning (Ye et al., 2022) addresses the manual schema-related challenges of traditional story generation methods, such as symbolic planning, by utilizing LLMs. By drawing upon common-sense knowledge extracted from these expansive language models, it's possible to recursively expand the SPG using a backward chaining approach from the goal event sentence, thus generating a consistent SPG. For our experiments, we employed this method, setting the last sentence of select stories from the ROCStories dataset as the goal event sentence and subsequently crafting the SPGs for experimentation. Consequently, stories generated through our skeleton selection will manifest as goal-driven narratives.

3 Skeleton Selection for SPG

The comprehensive algorithm for our proposed skeleton selection is presented in Section 3.1. The specifics of the event-based and graph-based approaches for selection are detailed in Section 3.2 and Section 3.3, respectively.

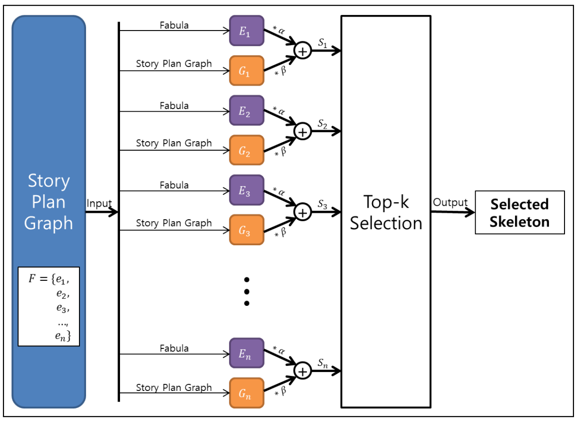

Figure 3: An illustration of the proposed algorithm. The skeleton selection score is computed for every event sentence in the fabula, and then the top k event sentences with the highest scores are selected to produce the skeleton.

3.1 Overview

As depicted in Figure 3, our skeleton selection algorithm computes the selection score for each event sentence e_i (where $1 \leq i \leq n$) within the fabula $F = \{e_1, e_2, ..., e_3\}$ as follows:

$$S_i(F) = \alpha E_i(F) + \beta G_i(F) \qquad (1)$$

where α represents the weight of the event-based score, and β signifies the weight of the graph-based score, with the constraint $\alpha + \beta = 1$ and $0 \leq \alpha, \beta \leq 1$. We aimed to adjust α to aptly blend the two scores. The overall process of computing the selection score is illustrated in Algorithm 1. Finally, the top-k event sentences were selected based on their selection scores. It's worth noting that the selected event sentences may not possess direct connections within the graph. For instance, in Figure 1, while "Ludo's work was taking a toll on his health" (denoted as e_1) and "Ludo got a prescription for the medicine from his doctor" (denoted as e_3) are selected, "Ludo drove himself to hospital" (denoted as e_2) may not be. Although e_1 and e_3 are not directly linked within the graph, readers can infer e_2 on their own, so there isn't an absolute necessity for e_2 to be selected alongside them.

Algorithm 1 Skeleton Selection for SPG

```
Input: plot created with plan_generation, plan graph's adj_list, vectorizer trained only with verbs, top_k
Initialize a fabula F ← plot
Initialize a goal G ← plot[-1]
tatal_vector ← vectorizer(connect all event in F)          ▷ vectorize only the verb tokens
each_vector ← vectorizer(F)
Initialize event_based_score ← {0,}          ▷ event-based score (In this paper, TF-IDF is used)
for each event in F do
    sum ← 0
    for e doach verb in event
        add total_vector(verb) * each_vector(verb) to sum
    end for
    add (event, sum) to event_based_score
end for
Initialize graph_based_score ← {0,}          ▷ graph-based score (In this paper, PageRank is used)
for each event in F do
    add (event, PageRank(adj_list, event) to graph_based_score
end for
selection_score ← α * event_based_score + β * graph_based_score          ▷ total score (α + β = 1)
top_k_selection ← sorted_and_pick(selection_score without G, top_k
return top_k_selection + G
```

3.2 Event-Based

The event-based score E_i is calculated based on the importance scores derived from the tf-idf of the event sentences within the fabula. Notably, condition sentences are anticipated to be represented in terms of causality with each event sentence during the computation of the graph-based score. Therefore, only the event sentences were utilized when calculating E_i. The computation of the event-based score E_i is as follows:

$$E_i(F) = \sum_{t \in e_i} idf(e_i, D) * tf(t, e_i) * idf(t, F) \quad (2)$$

where t denotes the events present in the event sentence, as highlighted in bold in Figure 4. The first term, $idf(e_i, D)$,references the inverse document frequency from the ROCStories dataset, D. The general inverse document frequency from a typical story dataset aids in filtering out mundane events. The second term, $tf(t, e_i)$, represents the term frequency and is employed to identify pivotal events within each event sentence. The final term, $idf(t, F)$, assists in filtering events that are commonly used locally. As observed in Figure 4, both "drive" and "get" appear with high frequency. If we do not filter them out through the final term, during the selection process, these terms might be disproportionately chosen, hindering the creation of an intriguing skeleton. To counteract this bias, we have incorporated the last term.

e_1) Ludo **drove** to his workplace in his car
e_2) Ludo **has completed** a new project that **needs** to **be completed** urgently
e_3) Ludo **got** the laptop from his company for work purposes
e_4) Ludo **was trying** to **impress** his boss by working hard
e_5) Ludo **got** the documents from his boss
e_6) Ludo was **working** long hours without **taking** enough breaks
e_7) Ludo's work **was taking** a toll on his health
e_8) Ludo **drove** himself to hospital
e_9) Ludo was not **feeling** well for a long time
e_{10}) Ludo **got** a prescription for the medicine from his doctor
e_{11}) Ludo **was recovering** from an illness
e_{12}) Ludo **drove** to the bank
e_{13}) Ludo **applied** for a credit card at his bank
e_{14}) Ludo **drove** back home from his workplace
e_{15}) Ludo **purchased** a subscription online **using** his credit card
e_{16}) Ludo **watched** a lot of movies on the subscription during the next week.

Figure 4: An example of a fabula. Events within the event sentences are highlighted in **bold**.

3.3 Graph-Based

Leveraging the information derived from the SPG, we assessed the significance of each event sentence node. Moreover, to mirror the importance of the causal relationships between event sentences and condition sentences, we employed the PageRank method to determine the graph-based score of each event sentence. In this context, we used the $distance(g, e_i)$ from the goal event sentence, g, to each node as a weight, accentuating events surrounding the goal. This approach was adopted due to our focus on selecting a skeleton for a goal-driven story. The graph-based score, G_i, is computed as follows:

$$G_i(F) = PageRank(adj\_list, distance(g, e_i)) \tag{3}$$

where adj_list represents the adjacency list of the SPG, and $distance$ is defined as the shortest path between g and e_i when at least one path exists between them, as described in Harary (1969). Notably, since every node in the SPG is generated through backward chaining from g, there are no instances where g and e_i are not connected.

4 Experiment

In this section, we evaluate our skeleton selection method using the SPGs generated based on the stories in ROCStories dataset. We first introduce the dataset, baseline, and evaluation methodology. Subsequently, we present the results in comparison with the baseline and discuss the implications of these findings.

4.1 Dataset

We employed the recently-introduced story planning method, Neural Story Planning, to generate the SPGs. For the goal event sentence, we utilized the final sentence from the stories in ROCStories dataset. Out of the generated plan graphs, we conducted experiments using 135 SPGs that adhered to the criteria of a fabula rather than a possible world. Each fabula comprises more than 15 event sentences.

4.2 Baseline

We used the LLM, GPT-3.5[1], to generate a skeleton for our baseline. By prompting, we provided the adjacency list of the SPG and the fabula, instructing it to select k event sentences, including the goal event sentence. Given that our study is centered on goal-driven storytelling, the final event sentence must be the goal event sentence. On many occasions, GPT-3.5 not only performed skeleton selection but also undertook ordering, given its advanced language processing capabilities. Since we wanted to compare only the skeleton selection performance, we rearranged the skeleton produced by GPT-3.5 in the order of the fabula and compared it with the skeleton selected using our method. Examples of prompts utilized to guide GPT-3.5 in selecting skeletons from the fabula can be found in Appendix A.

4.3 Evaluation

To evaluate whether the selected skeletons are 1) intriguing, 2) logical, and 3) cohesive towards the goal, we compared the skeleton produced by GPT-3.5 (A) and the skeleton selected using our method (B) using the following three questions:

- Interestingness Question: Which story was more interesting?
- Logic Coherency Question: Which story had coherent flow between sentences?
- Topic Coherency Question: Which story had overall consistency in theme?

For each of the three questions, we collected responses 10 times each for A or B to evaluate which skeleton, A or B, was selected more effectively. The responses were gathered using the GPT-3.5 version[2], which served as our baseline. For this evaluation, we set $\alpha = 0.5$ and $k = 10$. Examples of the prompts used for evaluation can be found in Appendix B.

Question Type	GPT-3.5 (%)	Ours (%)
Interestingness	28.89	**71.11**
Logic Coherency	21.48	**78.52**
Topic Coherency	11.85	**88.15**
average	20.84	**79.26**

Table 1: A/B test results for each question type at $\alpha = 0.5$. In the evaluation, items receiving a higher selection rate are highlighted in **bold**.

Question Type	$\alpha = 0$ (%)	$\alpha = 1$ (%)
Interestingness	64.89	**76.30**
Logic Coherency	**85.93**	74.04
Topic Coherency	**85.19**	83.70

Table 2: Proportion of selections favoring ours in the A/B test across question types, based on varying α. In the evaluation, items receiving a higher selection rate are highlighted in **bold**.

[1] 'gpt-3.5-turbo-16k-0613' version was used through OpenAI API. We opted for a specific version rather than the latest version to ensure consistency in our experiments.

[2] 'gpt-3.5-turbo-16k-0613'

Question Type	Ours (%)	simple (%)	no weight (%)
Interestingness	**71.11**	63.33	60.91
Logic Coherency	**78.52**	69.48	70.15
Topic Coherency	**88.15**	78.44	76.30
average	**79.26**	70.42	69.12

Table 3: Results from the ablation study evaluated at $\alpha = 0.5$. Here, 'simple' refers to the event-based method calculated using a straightforward tf-idf computation, while 'no weight' represents the graph-based method employing PageRank without any weighting. In the evaluation, items receiving a higher selection rate are highlighted in **bold**.

4.4 Results and Discussion

The results are presented in Table 1. Across all three question types, the skeleton selected using our method demonstrated a higher preference than the skeleton generated by GPT-3.5. Although these findings are based on evaluations from a LLM rather than human judgments, numerous prior studies (Xu et al., 2018; Ye et al., 2022; Chen et al., 2023) have utilized LLMs for auto-evaluation. Hence, it can be inferred that our algorithm performed a more effective skeleton selection.

To validate the efficacy of our proposed event-based and graph-based approaches, we assessed skeletons generated by adjusting the value of α. According to Equation 1, when $\alpha = 1$, the skeleton is selected solely based on the event-based method, and when $\alpha = 0$, it is based entirely on the graph-based method. The results are presented in Table 2. As we hypothesized, the graph-based only selection method more adeptly chose skeletons that were logical and coherent towards the goal. Additionally, the event-based approach seemed to aid in selecting more engaging skeletons. To further discern the utility of our proposed methods, we conducted an ablation study, as detailed in Section 4.5.

4.5 Ablation Study

To determine the impact of our proposed event-based score E_i and graph-based score G_i on the quality of skeleton selection, we conducted evaluations using a simple tf-idf and a PageRank that doesn't use weights, respectively. The results are displayed in Table 3. Across all question types, the skeleton selection method we proposed demonstrates superior performance. This suggests that both E_i and G_i which we proposed have been effectively applied in the skeleton selection process.

5 Conclusion

In this paper, we propose an algorithm to generate a narrative story skeleton by selecting the fabula using a SPG. Our approach employs both the event-based scheme, designed to emphasize pivotal event sentences in accordance with the story's funny, and the graph-based paradigm that emphasizes logical coherence and unity of event sentences within the story's structure. Collectively, they ascertain the overarching event sentences throughout the fabula. We employ the state-of-the-art, high-performing LLM, GPT-3.5, to auto-evaluate the interest, logical coherence, and unity of the skeleton. Our evaluations demonstrate the superior performance of our skeleton selection. Additionally, through an ablation study, we validated the efficacy of our proposed approach. This study's Skeleton Selection technique paves the way for further research tasks aiming to produce fully fleshed out narrative stories in an open-world setting, encompassing details and dialogues.

References

H Porter Abbott. 2002. Narrative and life. *The Cambridge Introduction to Narrative*, pages 1–12.

Mieke Bal and Christine Van Boheemen. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

David Bordwell. 1980. The musical analogy. *Yale French Studies*, (60):141–156.

Seymour Benjamin Chatman. 1978. *Story and discourse: Narrative structure in fiction and film*. Cornell university press.

Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2023. Storyer: Automatic story evaluation via ranking, rating and reasoning. *Journal of Natural Language Processing*, 30(1):243–249.

Gérard Genette. 1983. *Narrative discourse: An essay in method*, volume 3. Cornell University Press.

Arthur C Graesser, Kathy L Lang, and Richard M Roberts. 1991. Question answering in the context of stories. *Journal of Experimental Psychology: General*, 120(3):254.

F Harary. 1969. Graph theory addison-wesley reading ma usa.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Michael Lebowitz. 1985. Story-telling as planning and learning. *Poetics*, 14(6):483–502.

James R Meehan. 1977. Tale-spin, an interactive program that writes stories. In *Ijcai*, volume 77, pages 91–98.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The pagerank citation ranking: Bring order to the web. Technical report, Technical report, stanford University.

Julie Porteous and Marc Cavazza. 2009. Controlling narrative generation with planning trajectories: the role of constraints. In *Interactive Storytelling: Second Joint International Conference on Interactive Digital Storytelling, ICIDS 2009, Guimarães, Portugal, December 9-11, 2009. Proceedings 2*, pages 234–245. Springer.

Alec Radford and Jeffrey Wu. 2019. Rewon child, david luan, dario amodei, and ilya sutskever. 2019. *Language models are unsupervised multitask learners. OpenAI blog*, 1(8):9.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

Shlomith Rimmon-Kenan. 2002. Towards... afterthoughts, almost twenty years later. *SR-K.: Narrative Fiction. Contemporary Poetics. 2nd ed. London and New York: Routledge*, pages 134–149.

Marie-Laure Ryan. 1991. *Possible worlds, artificial intelligence, and narrative theory*. Indiana University Press.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Richard Walsh. 2001. Fabula and fictionality in narrative theory. *Style*, 35(4):592–606.

Stephen Ware and R Young. 2011. Cpocl: A narrative planner supporting conflict. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 7, pages 97–102.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. *arXiv preprint arXiv:1808.06945*.

Anbang Ye, Christopher Cui, Taiwei Shi, and Mark O Riedl. 2022. Neural story planning. *arXiv preprint arXiv:2212.08718*.

A Prompts for the skeleton selection

Examples of the prompts utilized to guide GPT-3.5 in performing skeleton selection are provided below:

role =

> You create a skeleton story by selecting events from the tree-structured story planner. You have to look at the story planner given an adjacency list and choose 9 events in event list. The criteria for selecting events can be freely defined. Please select an appropriate event considering the fun of the event, causal rink, goal sentence, etc.

content =

> goal: Ludo watched a lot of movies on the subscription during the next week.
>
> adjacency list:
> Ludo watched a lot of movies on the subscription during the next week.:set()
> I; A subscription for watching movies:Ludo watched a lot of movies on the subscription during the next week.
> Ludo purchased a subscription online using his credit card:I; A subscription for watching movies
> ...

event list:
Ludo drove to his workplace in his car
Ludo has completed a new project that needs to be completed urgently
Ludo got the laptop from his company for work purposes
Ludo was trying to impress his boss by working hard
...

Question: Choose 9 events in event list.
Answer:

B Prompts for A/B testing

Examples of the prompts employed to facilitate the A/B testing are presented below:

role =

You are the story evaluator. You just have to look at Story A and Story B, and answer the questions only with "A" or "B".

content =

Story A:
(GPT-3.5's skeleton created with Appendix A)
...

Story B
(skeleton selected with our method)
...

Question: Which story was more interesting?
Answer: