Leveraging LLM-Constructed Graphs for Effective Goal-Driven Storytelling

Taewoo Yoo[1,*,†], Yun-Gyung Cheong[1,†]

[1]*SungKyunKwan University (SKKU) / Suwon, South Korea*

Abstract

While advanced language models, such as Large Language Models (LLMs), have demonstrated potential in generating various types of text, including narratives, they often struggle to maintain semantic consistency. In narrative theory, skeleton selection refers to deriving a story's backbone by choosing only the pivotal events, or nucleus, from the comprehensive story world (fabula), ensuring a focused and coherent narrative structure. To address the challenges faced by LLMs, we utilize Story Plan Graphs (SPGs)—a form of Knowledge Graphs—to ensure logical soundness for skeleton construction. When evaluated against GPT-3.5 using the ROCStories dataset, our approach demonstrates enhanced skeleton selection capabilities, offering an efficient solution for storytelling.

Keywords

LLMs (Large Language Models), SPGs (Story Plan Graphs), KGs (Knowledge Graphs), Narrative generation, Goal-driven storytelling

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. A story refers to a series of events linked by causality, experienced or enacted by actors [1, 2, 3]. For instance, "Mary woke up late. She missed the bus to work. Her boss was unhappy" is considered a story, whereas "Mary woke up late. She wore a blue dress to work. The coffee machine was broken." is non-narrative.

A coherent and engaging story demands that each sentence logically follows the previous one. This means that the events, actions, and dialogue in the story must be linked by cause and effect, ensuring that the overall narrative makes sense for the reader. Furthermore, crafting a story that engages and entertains the reader presents a notable challenge. Consequently, narrative generation has captured the interest of researchers for decades and has become a topic of intensive investigation with the advent of LLMs enabled by Transformer-based language models [4, 5].

While LLMs offer significant improvements in narrative generation, they still face challenges in maintaining deep semantic coherence, avoiding repetition, and producing highly specific and creative responses [6]. Moreover, LLMs sometimes exhibit reasoning errors and inconsistent responses due to the lack of an underlying belief system and the reliance on probabilistic patterns from its training data [7, 8].

To address these limitations, utilizing knowledge graphs (KGs) and reasoning frameworks can be a solution. Traditional solutions to story generation, such as symbolic planning [9, 10, 11, 12, 13], can infer causal relationships between events. Additionally, it offers a means to model semantic dependencies in the form of a graph. This is conceptually similar to KGs, which represent information in a structured format. Thus, this paper presents Story Plan Graph (SPGs) as a form of Knowledge Graph, specifically tailored for narrative story generation.

The First international OpenKG Workshop: Large Knowledge-Enhanced Models, August 03, 2024, Jeju Island, South Korea

[*] Corresponding author.

[†] These authors contributed equally.

woo990307@skku.edu (T. Yoo); aimecca@skku.edu (Y. Cheong)

https://github.com/mzyt3569 (T. Yoo); https://inglab.github.io/ (Y. Cheong)

0009-0005-2501-7620 (T. Yoo)

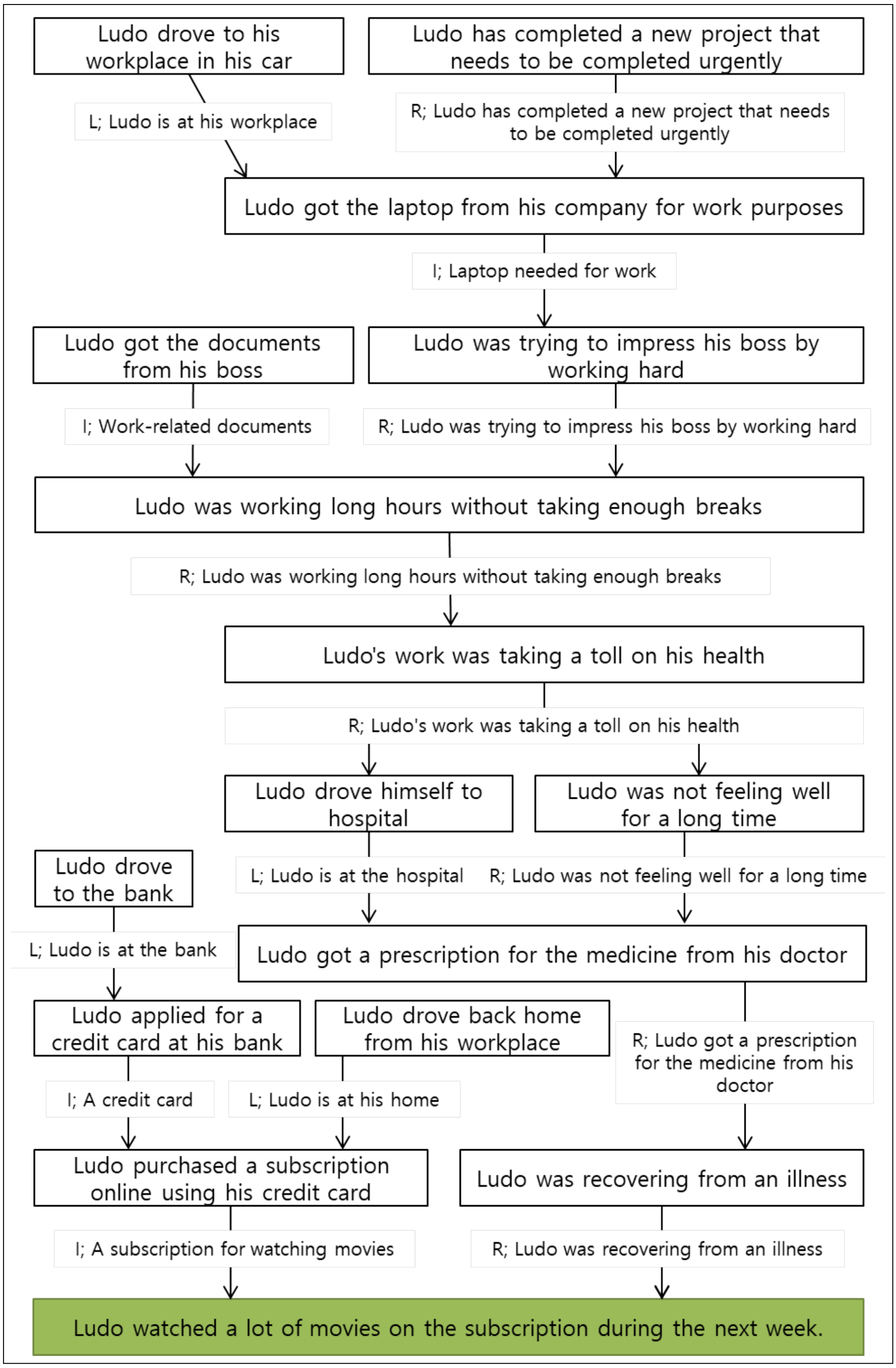
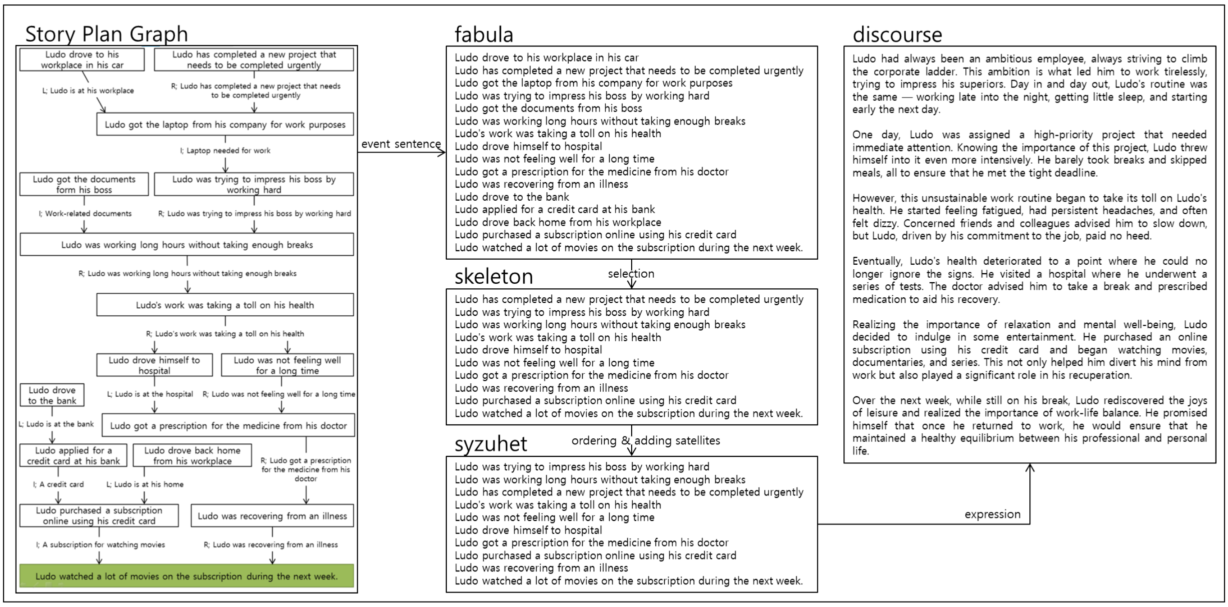

Ludo drove to his workplace in his car
Ludo has completed a new project that needs to be completed urgently
L; Ludo is at his workplace
R; Ludo has completed a new project that needs to be completed urgently
Ludo got the laptop from his company for work purposes
I; Laptop needed for work
Ludo got the documents from his boss
Ludo was trying to impress his boss by working hard
I; Work-related documents
R; Ludo was trying to impress his boss by working hard
Ludo was working long hours without taking enough breaks
R; Ludo was working long hours without taking enough breaks
Ludo's work was taking a toll on his health
R; Ludo's work was taking a toll on his health
Ludo drove himself to hospital
Ludo was not feeling well for a long time
Ludo drove to the bank
L; Ludo is at the hospital
R; Ludo was not feeling well for a long time
L; Ludo is at the bank
Ludo got a prescription for the medicine from his doctor
Ludo applied for a credit card at his bank
Ludo drove back home from his workplace
R; Ludo got a prescription for the medicine from his doctor
I; A credit card
L; Ludo is at his home
Ludo purchased a subscription online using his credit card
Ludo was recovering from an illness
I; A subscription for watching movies
R; Ludo was recovering from an illness
Ludo watched a lot of movies on the subscription during the next week.

Figure 1: An illustrative example of a SPG generated via a planning algorithm. 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 denote condition sentences. Within the condition sentences, 'I' stands for item need condition, 'L' represents location condition, and 'R' indicates reason condition.

A story can be analyzed via a tripartite model, which include the notions of *fabula*, *syuzhet*, and *discourse* [14, 15, 16, 17]. The term fabula provides the raw content of a story, the syuzhet selects and organizes that content, and the discourse presents it to the audience.


Story Plan Graph
Ludo drove to his workplace in his car
Ludo has completed a new project that needs to be completed urgently
L; Ludo is at his workplace
R; Ludo has completed a new project that needs to be completed urgently
Ludo got the laptop from his company for work purposes
I; Laptop needed for work
Ludo got the documents form his boss
Ludo was trying to impress his boss by working hard
I; Work-related documents
R; Ludo was trying to impress his boss by working hard
Ludo was working long hours without taking enough breaks
R; Ludo was working long hours without taking enough breaks
Ludo's work was taking a toll on his health
R; Ludo's work was taking a toll on his health
Ludo drove himself to hospital
Ludo was not feeling well for a long time
Ludo drove to the bank
L; Ludo is at the hospital
R; Ludo was not feeling well for a long time
L; Ludo is at the bank
Ludo got a prescription for the medicine from his doctor
Ludo applied for a credit card at his bank
Ludo drove back home from his workplace
R; Ludo got a prescription for the medicine from his doctor
I; A credit card
L; Ludo is at his home
Ludo purchased a subscription online using his credit card
Ludo was recovering from an illness
I; A subscription for watching movies
R; Ludo was recovering from an illness
Ludo watched a lot of movies on the subscription during the next week.
event sentence
fabula
selection
skeleton
ordering & adding satellites
syuzhet
expression
discourse

fabula

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
Ludo got the documents from his boss
Ludo was working long hours without taking enough breaks
Ludo's work was taking a toll on his health
Ludo drove himself to hospital
Ludo was not feeling well for a long time
Ludo got a prescription for the medicine from his doctor
Ludo was recovering from an illness
Ludo drove to the bank
Ludo applied for a credit card at his bank
Ludo drove back home from his workplace
Ludo purchased a subscription online using his credit card
Ludo watched a lot of movies on the subscription during the next week.

skeleton

Ludo has completed a new project that needs to be completed urgently
Ludo was trying to impress his boss by working hard
Ludo was working long hours without taking enough breaks
Ludo's work was taking a toll on his health
Ludo drove himself to hospital
Ludo was not feeling well for a long time
Ludo got a prescription for the medicine from his doctor
Ludo was recovering from an illness
Ludo purchased a subscription online using his credit card
Ludo watched a lot of movies on the subscription during the next week.

syuzhet

Ludo was trying to impress his boss by working hard
Ludo was working long hours without taking enough breaks
Ludo has completed a new project that needs to be completed urgently
Ludo's work was taking a toll on his health
Ludo was not feeling well for a long time
Ludo drove himself to hospital
Ludo got a prescription for the medicine from his doctor
Ludo purchased a subscription online using his credit card
Ludo was recovering from an illness
Ludo watched a lot of movies on the subscription during the next week.

discourse

Ludo had always been an ambitious employee, always striving to climb the corporate ladder. This ambition is what led him to work tirelessly, trying to impress his superiors. Day in and day out, Ludo's routine was the same — working late into the night, getting little sleep, and starting early the next day.

One day, Ludo was assigned a high-priority project that needed immediate attention. Knowing the importance of this project, Ludo threw himself into it even more intensively. He barely took breaks and skipped meals, all to ensure that he met the tight deadline.

However, this unsustainable work routine began to take its toll on Ludo's health. He started feeling fatigued, had persistent headaches, and often felt dizzy. Concerned friends and colleagues advised him to slow down, but Ludo, driven by his commitment to the job, paid no heed.

Eventually, Ludo's health deteriorated to a point where he could no longer ignore the signs. He visited a hospital where he underwent a series of tests. The doctor advised him to take a break and prescribed medication to aid his recovery.

Realizing the importance of relaxation and mental well-being, Ludo decided to indulge in some entertainment. He purchased an online subscription using his credit card and began watching movies, documentaries, and series. This not only helped him divert his mind from work but also played a significant role in his recuperation.

Over the next week, while still on his break, Ludo rediscovered the joys of leisure and realized the importance of work-life balance. He promised himself that once he returned to work, he would ensure that he maintained a healthy equilibrium between his professional and personal life.

Figure 2: Depicts the process of generating a discourse from an SPG. This paper details the process of generating a skeleton from an SPG.

Drawn upon this narrative analysis theory, we aim to construct a narrative as a story plan graph (SPG) at the fabula layer and select core events as a skeleton at the syuzhet layer. Specifically, this study investigates which events should be chosen from the modeled SPG to construct the most effective skeleton in terms of coherency, logicality, and interestingness.

The key contributions of this research are enumerated as follows:

- We propose a new method leveraging story plan knowledge graphs to construct a coherent story.
- We propose an efficient content selection procedure based on the well-established significance metrics, TF-IDF [18] and the PageRank [19] algorithm.
- We conducted an automated evaluation utilizing GPT-3.5 and the ROCStories dataset [20]. The results indicate that our approach effectively constructs the skeleton, by accurately identifying and prioritizing key events within the narrative, ensuring both relevance and coherence in the given story.

In this study, we examine how symbolic knowledge, specifically Story Plan Graphs (SPGs), and algorithms can enhance narrative generation using Large Language Models (LLMs).

The structure of the paper is as follows: Section 2 reviews related works; Section 3 describes our skeleton selection approach; Section 4 presents the experiment and discusses the results; and finally, Section 5 concludes with future work.

2. Background and Related Work

2.1. Narrative Analysis Theory

The employment of the bipartite model—story and discourse—in analyzing narrative has a long history in narratology [14]. In this model, story refers to the content plane of narrative whereas discourse represents its expression plane.

Some narrative theorists [15, 16, 17] maintain that different stories emerging from the same story material is rooted in the existence of an abstract entity called the narrator, who decides what to tell and when to tell it. To distinguish the narrator's role from the discourse, they propose a three-tiered model of narrative consisting of the fabula, the sjuzhet, and the narrative discourse. 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*' [21], 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 of the story–named *nucleus* [14]. 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 [14]. The '*discourse*' represents the syuzhet as expressed through mediums like text or film. Our research focuses on skeleton selection, grounded in the aforementioned theories and definitions.

2.2. Computational Approaches to Story Generation

Traditional story generation systems leverage symbolic approaches such as inference and planning algorithms. These systems are divided into author-centric and character-centeric approaches. Talespin [9] generates stories by modeling the goals and actions of characters and constructing narratives through their interactions. Universe [22] is a system focused on the creative aspects of storytelling, designed as an aid for writers. It synthesizes various story elements into a plot through interaction with humans.

Minstrel [23] is a knowledge-based system for storytelling, emphasizing character and plot development. It simulates creative problem-solving in story generation, employing methods to use existing knowledge in novel ways. Mexica [24] aims to model the creative process of story writing, specifically creating narratives about the lives of early Mexican natives. Its approach emphasizes creativity and emotional connections to deepen the narrative generated.

Fabulist [25] is a story generation architecture that models story structure and character intentions, considering the causes and consequences of events to create narratives. Virtual StoryTeller [26] employs a multi-agent approach to generate stories. Each agent, with its independent knowledge and goals, interacts in the story development process, selecting actions that contribute to story creation. Our work references these methodologies to study skeleton selection methods.

Existing story generation models often struggle to maintain consistency. Various approaches have been researched to address this issue. For instance, Xie *et al.* [27] investigated whether large pre-trained language models could learn storytelling with few examples. Additionally, Peng *et al.* [28] proposed a method to improve the consistency and thematic coherence of neural network-based story generation using reader models. Furthermore, Wang *et al.* [29] conducted a comprehensive survey on open-world story generation with structured knowledge enhancement, exploring ways to improve the logical coherence of generated stories. Xu *et al.* [30] proposed the MEGATRON-CNTRL framework, which integrates an external knowledge base to enable controllable story generation. These studies illustrate additional methods for LLMs to maintain coherent and logical story structures beyond simple text generation capabilities. In this research, we utilized SPGs—a form of KGs—to enhance the consistent storytelling abilities of LLMs.

Neural Story Planning [31] 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 is possible to recursively expand the SPG using a backward chaining approach from the goal event sentence, thus generating a consistent SPG. In this paper, we leverage these SPGs as a form of KGs, integrating the structured, logical representation of symbolic planning with the language-based knowledge generated by an LLM.

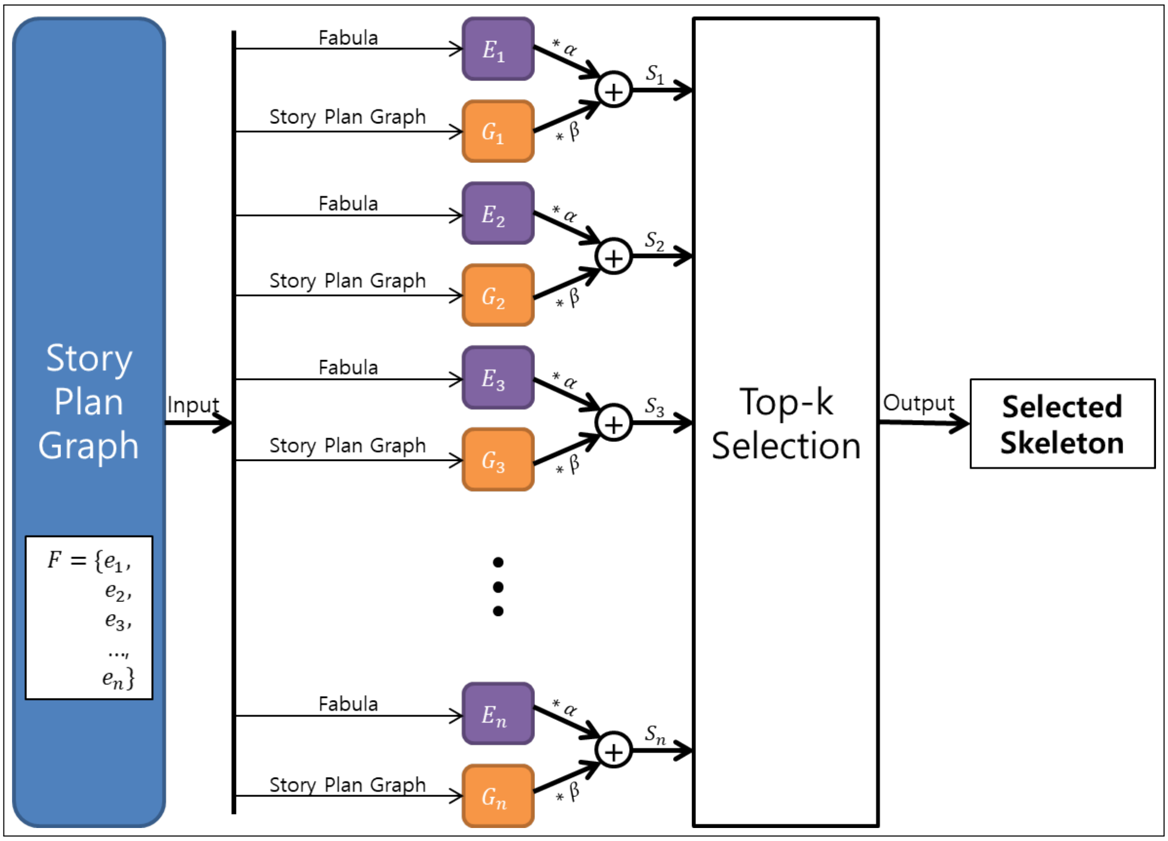


Figure 3: An illustration of the proposed algorithm. The skeleton selection score is computed for every event sentence in the fabula, and then the top k event sentences with the highest scores are selected to produce the skeleton.

3. Skeleton Selection

3.1. Overview

First, we construct the SPGs employing the Neural Story Planning method [31], setting the last sentence of select stories from the ROCStories dataset as the goal event sentence and subsequently constructing the SPGs.

Figure 3 depicts our skeleton selection algorithm, which 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:

$$S_i(F) = \alpha E_i(F) + \beta G_i(F) \tag{1}$$

where α represents the weight of the event-based score, and β denotes the weight of the graph-based score, with the constraint $\beta = 1 - \alpha$ and $0 \leq \alpha, \beta \leq 1$. We aim to adjust α to blend the two scores.

The overall process of computing the selection score follows the steps as shown in Algorithm 1:

1. **Initialize a fabula F:** This step initializes the entire story plot.
2. **Initialize a goal G:** The last sentence of the story is set as the goal event.
3. **Vectorize Events:** All events in the plot are vectorized to evaluate their importance.
4. **Compute Event-Based Score:** The event-based score for each event sentence is calculated using tf-idf.
5. **Compute Graph-Based Score:** The graph-based score is calculated using the PageRank algorithm and the distance from the goal event.
6. **Combine Scores:** The event-based and graph-based scores are combined to compute the final selection score.

Algorithm 1 Skeleton Selection for SPG

```
Input: Story Plan Graph
Parameter: plot created with plan_generation, plan graph's adj_list, vectorizer trained only with verbs, top_k
Output: Selected Skeleton
Initialize a fabula $F$ ← plot
Initialize a goal $G$ ← plot[-1]
total_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 each verb in event do
        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, distance($G$, event))) to graph_based_score
end for
selection_score ← $\alpha$ * event_based_score + $\beta$ * graph_based_score {total score ($\alpha + \beta = 1$)}
top_k_selection ← sorted_and_pick(selection_score without $G$, top_k
return top_k_selection + $G$
```

7. **Select Top-k Events:** The top-k event sentences are selected based on their final selection scores.

This algorithm ensures logical coherence and an interesting story composition by evaluating the causal relationships between event sentences (e_i) generated through backward chaining from the goal event sentence (g) in the SPG.

Finally, the top-k event sentences are selected based on their selection scores. Please note that the selected event sentences may not be directly linked 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. Thus, readers can comprehend the story without e_2 being selected.

3.2. Event-Based Score

The event-based score E_i is calculated based on the importance derived from the tf-idf of the event sentences within the fabula. Condition sentences appear associated with causal links between event sentences 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:

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

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 aids in filtering out the events that occur frequently throughout the datasets. 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 frequently used locally. For instance, in the story shown in Figure 4, both the words ‘drive’ and ‘get’ appear frequently. The final term helps reduce the probability of selecting these commonly occurring words.

3.3. Graph-Based Score

Leveraging the information derived from the SPG, we assess the significance of each event sentence node. To incorporate the importance of the causal relationships between event sentences and condition sentences, we employ the PageRank method to determine the graph-based score of each event sentence.

We also consider the $distance(g, e_i)$ from the goal event sentence g to each node as a weight, emphasizing events surrounding the goal. This approach was chosen to align with our focus on selecting a skeleton for a goal-driven story. The graph-based score, G_i, is computed using the following equation:

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

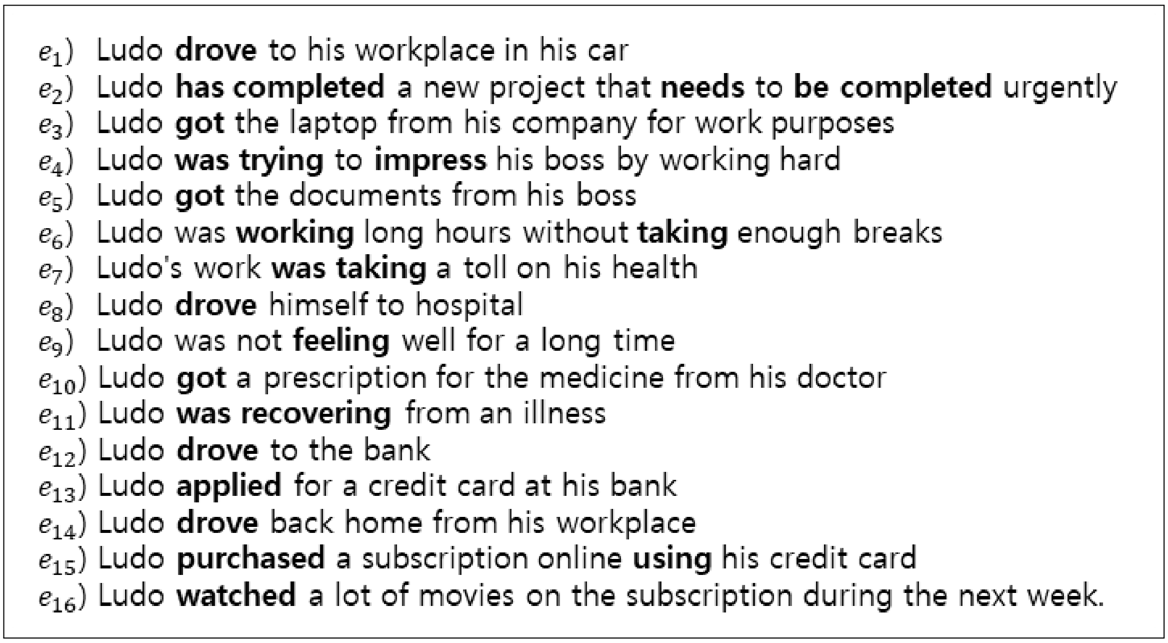
where adj_list represents the adjacency list of the SPG, and $distance$ is defined as the shortest path between g and e_i when at least one path exists between them, as described in Harary [32]. Notably, since every node in the SPG is generated through backward chaining from g, there are no instances 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 describe the dataset, baseline, and evaluation methodology. Subsequently, we present the results in comparison with the baseline and ablation study, discussing 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 the ROCStories



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**.

dataset. The ROCStories dataset, introduced by Mostafazadeh et al. [20], is a collection of 100,000 short commonsense stories designed for research in commonsense reasoning and story understanding.

Each story in the ROCStories dataset consists of five sentences that describe everyday scenarios, providing a rich source of diverse narrative structures. This makes it particularly suitable for evaluating story generation and skeleton selection methods.

From 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. The selection criteria ensured that the stories used in our experiments maintained a level of complexity suitable for testing our skeleton selection algorithm.

Additionally, the ROCStories dataset allows for the testing of narrative coherence and logical progression, as the stories inherently contain causal links and event dependencies. This characteristic of the dataset was crucial for evaluating the effectiveness of our Story Plan Graph-based approach to skeleton selection.

4.2. Baseline

We used GPT-3.5[1] to generate a skeleton for our baseline. We provided the adjacency list of the SPG and the fabula through prompting, instructing it to select k event sentences, including the goal event sentence. Given that our study focuses on goal-driven storytelling, the final event sentence represents the goal event. We noted that GPT-3.5 not only performed skeleton selection but also undertook ordering. Since we need to compare only the skeleton selection performance, we rearrange the skeleton produced by GPT-3.5 to match the order of the fabula. Examples of prompts designed to guide GPT-3.5 in selecting skeletons from the fabula are as follows:

> **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
> ...

[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.

Question: Choose 9 events in event list.
Answer:

We additionally conducted skeleton selection using ChatGPT[2]. By comparing the skeleton selection performance with ChatGPT, known for its high proficiency in a wide range of linguistic tasks, we aimed to assess the effectiveness of our selection algorithm. We utilized ChatGPT 4 with the same prompts used in GPT-3.5 for interactive tasks.

4.3. Evaluation Method

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: Which story was more interesting?
- Logic Coherency: Which story had coherent flow between sentences?
- Topic Coherency: 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[3], which served as our baseline. For this evaluation, we set $\alpha = 0.5$ and $k = 10$.

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:
1. Horace was transported to a location where he can freely move around by walking
2. Horace hailed a taxi on the street
3. Horace took a taxi to the car dealership
4. Horace bought a car from a dealership
5. Horace drove his car to the hardware store
6. Horace had been using the lightbulb in his bathroom for a long time until it burned out
7. The old lightbulb burned out after being used for a long time
8. Horace asked a store employee for assistance
9. Horace bought a new lightbulb from a hardware store
10. Horace is glad the lightbulb in his bathroom is no longer dead.

Story B:
1. Horace didn’t have a choice in inheriting his functional legs
2. Horace inherited his pair of functional legs from his parents
3. Horace has had the ability to walk since he was born
4. Horace walked to the street where he hailed the taxi
5. Horace hailed a taxi on the street
6. Horace had been using the lightbulb in his bathroom for a long time until it burned out
7. The old lightbulb burned out after being used for a long time
8. Horace had a doubt
9. Horace bought a new lightbulb from a hardware store
10. Horace is glad the lightbulb in his bathroom is no longer dead.

[2] https://chat.openai.com/
[3] ‘gpt-3.5-turbo-16k-0613’

Question: Which story was more interesting?
Answer:

In this example, Story A is skeleton selected with GPT-3.5, and Story B as skeleton selected with our method. The order of Story A and Story B is determined randomly. The rationale for randomizing the order in our evaluations stems from the positional bias found in large language models, as identified in recent research [33]. To mitigate this bias, we randomized the story sequence and accordingly structured our prompts.

4.4. Results and Discussion

As presented in Table 1, the skeleton selected using our method was favored over the skeleton generated by GPT-3.5 across all three question types. Although these findings are based on evaluations by an LLM rather than human judgments, numerous prior studies [31, 33] have utilized LLMs for auto-evaluation. Hence, it can be inferred that our algorithm performed a more effective skeleton selection.

Table 2 report the results from the experiments comparing our skeleton selection method with that of ChatGPT. The preference for our algorithm, though marginally higher, indicates that our algorithm can exhibit comparable performance in the skeleton selection task to the commercial large-scale language models.

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 3. 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.

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	ChatGPT (%)	Ours (%)
Interestingness	48.15	**51.85**
Logic Coherency	45.93	**54.07**
Topic Coherency	42.96	**57.04**
average	45.68	**54.32**

Table 2
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 3
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**.

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 4
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**.

4.5. 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 4.

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 important events from the fabula using a Story Plan Graph (SPG), which emphasizes the logical coherence of event sentences within the story's structure. Our approach also considers an event-based scheme to include pivotal events based on their occurrences in the fabula. Collectively, these methods ensure the inclusion of overarching event sentences throughout the fabula.

We employ GPT-3.5 to automatically evaluate the interest, logical coherence, and unity of the skeleton. The results demonstrate that our skeleton selection algorithm outperforms GPT-3.5 and shows comparable performance to ChatGPT, while offering greater efficiency in terms of API usage fees and physical resources.

We plan to conduct a comprehensive formal evaluation using state-of-the-art LLMs to validate the efficacy of our proposed approach. By integrating SPGs as a form of Knowledge Graph and LLMs, we believe that this paper contributes to the computational storytelling community by combining the strengths of symbolic and neural methods for reliable knowledge processing.

Acknowledgments

This work was partly supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MIST) (No.RS-2024-00357849), the Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No.RS-2024-00413839), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00421, Artificial Intelligence Graduate School Program (Sungkyunkwan University)).

References

[1] D. Bordwell, The musical analogy, Yale French Studies (1980) 141–156.

[2] M. Bal, C. Van Boheemen, Narratology: Introduction to the theory of narrative, University of Toronto Press, 2009.

[3] H. P. Abbott, Narrative and life, The Cambridge Introduction to Narrative (2002) 1–12.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[5] A. Radford, J. Wu, Rewon child, david luan, dario amodei, and ilya sutskever. 2019, Language models are unsupervised multitask learners. OpenAI blog 1 (2019) 9.

[6] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, arXiv preprint arXiv:1904.09751 (2019).

[7] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report, 2024. arXiv:2303.08774.

[8] J. A. Baktash, M. Dawodi, Gpt-4: A review on advancements and opportunities in natural language processing, 2023. arXiv:2305.03195.

[9] J. R. Meehan, Tale-spin, an interactive program that writes stories., in: Ijcai, volume 77, 1977, pp. 91–98.

[10] M. Lebowitz, Story-telling as planning and learning, Poetics 14 (1985) 483–502.

[11] J. Porteous, M. Cavazza, 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, Springer, 2009, pp. 234–245.

[12] M. O. Riedl, R. M. Young, Narrative planning: Balancing plot and character, Journal of Artificial Intelligence Research 39 (2010) 217–268.

[13] S. Ware, R. Young, Cpocl: A narrative planner supporting conflict, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 7, 2011, pp. 97–102.

[14] S. B. Chatman, Story and discourse: Narrative structure in fiction and film, Cornell university press, 1978.

[15] G. Genette, Narrative discourse: An essay in method, volume 3, Cornell University Press, 1983.

[16] S. Rimmon-Kenan, Towards... afterthoughts, almost twenty years later, SR-K.: Narrative Fiction. Contemporary Poetics. 2nd ed. London and New York: Routledge (2002) 134–149.

[17] R. Walsh, Fabula and fictionality in narrative theory, Style 35 (2001) 592–606.

[18] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of documentation 28 (1972) 11–21.

[19] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bring order to the web, Technical Report, Technical report, stanford University, 1998.

[20] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, J. Allen, A corpus and evaluation framework for deeper understanding of commonsense stories, arXiv preprint arXiv:1604.01696 (2016).

[21] M.-L. Ryan, Possible worlds, artificial intelligence, and narrative theory, Indiana University Press, 1991.

[22] M. Lebowitz, Creating a story-telling universe (1983).

[23] S. R. Turner, Minstrel: a computer model of creativity and storytelling, University of California, Los Angeles, 1993.

[24] R. P. y Pérez, MEXICA: a computer model of creativity in writing, Norman Spinrad, 1999.

[25] M. O. Riedl, R. M. Young, A planning approach to story generation for history education, in: Proceedings of the 3rd International Conference on Narrative and Interactive Learning Environments, 2004, pp. 41–48.

[26] M. Theune, S. Faas, A. Nijholt, D. Heylen, The virtual storyteller: Story creation by intelligent agents, in: Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference, volume 204215, 2003, p. 116.

[27] Z. Xie, T. Cohn, J. H. Lau, The next chapter: A study of large language models in storytelling, 2023. arXiv:2301.09790.

[28] X. Peng, K. Xie, A. Alabdulkarim, H. Kayam, S. Dani, M. O. Riedl, Guiding neural story generation with reader models, 2022. arXiv:2112.08596.

[29] Y. Wang, J. Lin, Z. Yu, W. Hu, B. F. Karlsson, Open-world story generation with structured knowledge enhancement: A comprehensive survey, 2023. arXiv:2212.04634.

[30] P. Xu, M. Patwary, M. Shoeybi, R. Puri, P. Fung, A. Anandkumar, B. Catanzaro, Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models, 2020. arXiv:2010.00840.

[31] A. Ye, C. Cui, T. Shi, M. O. Riedl, Neural story planning, arXiv preprint arXiv:2212.08718 (2022).

[32] F. Harary, Graph theory addison-wesley reading ma usa, 1969.

[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, arXiv preprint arXiv:2306.05685 (2023).