Survey paper

Open-world story generation with structured knowledge enhancement: A comprehensive survey

Yuxin Wang [a,1], Jieru Lin [b], Zhiwei Yu [c], Wei Hu [a,*], Börje F. Karlsson [d,1]

[a] *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, Jiangsu, China*
[b] *Harbin Institute of Technology, Harbin, 150001, Heilongjiang, China*
[c] *Microsoft Research Asia, Beijing, 100080, Beijing, China*
[d] *Beijing Academy of Artificial Intelligence, 100086, Beijing, China*

ARTICLE INFO

ABSTRACT

Storytelling and narrative are fundamental to human experience, intertwined with our social and cultural engagement. As such, researchers have long attempted to create systems that can generate stories automatically. In recent years, powered by deep learning and massive data resources, automatic story generation has shown significant advances. However, considerable challenges, like the need for global coherence in generated stories, still hamper generative models from reaching the same storytelling ability as human narrators. To tackle these challenges, many studies seek to inject structured knowledge into the generation process, which is referred to as structured knowledge-enhanced story generation. Incorporating external knowledge can enhance the logical coherence among story events, achieve better knowledge grounding, and alleviate over-generalization and repetition problems in stories. This survey provides the latest and comprehensive review of this research field: (i) we present a systematic taxonomy regarding how existing methods integrate structured knowledge into story generation; (ii) we summarize involved story corpora, structured knowledge datasets, and evaluation metrics; (iii) we give multidimensional insights into the challenges of knowledge-enhanced story generation and cast light on promising directions for future study.

1. Introduction

Humans think in terms of stories, the world is understood in terms of stories, and people often approach problem-solving and new ideas by referencing stories they already understand [1]. Storytelling and narrative are therefore fundamental to human experience, intertwined with our social and cultural engagement [2]. Not only do people spread, preserve, and revise stories for informational, educational, or entertaining purposes [3–5], but the human brain has a natural affinity for creating them [6]. As such, *story generation* has consistently drawn attention from the natural language processing (NLP) community [7]. Beyond simple text generation, automatic story generation concerns spontaneously generating a coherent narrative, often prompted by a context or goal [8].

Early efforts on automatic story generation were generally based on formalisms or symbolic approaches [9–11]. Multiple storytelling systems were developed along these lines to assist authors focusing on specific scenarios or to enable interactive storytelling. Some sig-

nificant systems include: Tailspin [12] and Minstrel [13] among the pioneering systems, DEFACTO [14], Mimesis [15], IDtension [16], and Façade [17] for interactive drama, GADIN [18] and IDA [19] for later author-centric approaches, and Suspenser [20] for suspense novel writing. The majority of these works targeted *close-world* stories, where all the story elements, such as characters, emotions, places, and events, are constrained by pre-defined world settings. A significant example of formalism that inspired such approach is Propp's study of folktales [21], which summarizes the composition of tales and patterns of event development. Though these symbolism- and formalism-based methods can automatically generate stories with little human intervention, they require arduous knowledge engineering to build word settings for specific domains.

With the emergence of deep learning algorithms and the increased capabilities of generative models, efforts started to shift to neural story generation. These models can produce text sequences by modeling the probability distribution of word sequences (i.e. language model) [22]. This process requires no pre-defined world settings and is referred to

---

as *open-world* story generation. Later, based on the Transformer architecture [23], different efforts begin to build large pre-trained language models (PLMs) by feeding them with massive amounts of unsupervised data.

With such massive training data and huge parameter spaces, PLMs show a great ability to capture syntactic and semantic information within text [24]. They can produce grammatically correct and semantically intelligible short stories, even sometimes surprise readers. Recent trends on story generation tend to be based on PLMs and conduct specific manipulations to cater to different scenarios and requirements, such as persona control [25–27], content planning [28–31], human-AI collaboration [32–34], and explainability [35].

However, there are still many known problems and open issues in leveraging PLMs to generate stories. One major challenging issue is the lack of logical coherence and consistency [36,37], especially in long text. PLMs have shown unsatisfactory performance in capturing the semantic dependency among sentences and events. Generated events can be irrelevant, contradictory, and logically "out-of-order". The personas and emotions of protagonists can be "out-of-character"; and storylines can deviate from the theme or genre.

Another open problem is repetition [36,38]. PLMs are still likely to get stuck in a state which keeps producing repetitive token sequences. Moreover, PLMs show deficient knowledge grounding [39]. As a result, generated stories can be bland, not informative and specific.

Furthermore, story generation via PLMs lack controllability [40]. It is hard to control the content produced by PLMs beyond the given prompt, such as enforcing or editing the narrative style, character persona, specific events, or story topic.

To alleviate these problems, many recent studies begin to seek external knowledge to enhance story generation (introduced in Section 3). The intuition being that, by integrating structured knowledge that encodes explicit facts or rules into the generation process, such logical rules and inter-dependencies help to make generated stories more coherent and consistent. In addition, learning from knowledge-intensive data helps to achieve better knowledge grounding. Structured knowledge enhancement has shown to be effective in lifting the quality of generated stories [37,41,42].

There have been a few recent surveys on text and story generation in general. But these do not cover state-of-the-art developments in using structured knowledge to improve story generation. Hou et al. [43] give an overview of early deep learning-based works that train seq2seq models from scratch. They classify these works into three categories according to the user constraints: theme-oriented, storyline-oriented, and human-machine interaction-oriented. Alabdulkarim et al. [44] focus on the challenges of PLM-based story generation models and specific approaches to address them. They highlight the challenges of controllability and commonsense. Alhussain et al. [7] gives a more thorough development of story generation, especially early symbolism- and formalism-based works. Despite its somewhat comprehensive scope, it does not elaborate on recent structured knowledge-enhanced works nor presents a taxonomy. Lastly, Yu et al. [45] summarize the designs and techniques used in recent knowledge-enhanced text generation literature. However, they cover a broad spectrum including dialogue systems, question answering, and summarization, while lacking an organized and insightful discussion on story generation.

Recent years have witnessed big progress in structured knowledge-enhanced story generation in terms of scenario, methodology, data sources, and story quality measurements. A systematical and insightful overview of existing methods is necessary for advancing the field. Yet, there are still many challenges and bottlenecks requiring perceptive analysis and specific solutions.

Our survey serves a latest and comprehensive view of structured knowledge-enhance story generation to fill the gaps of previous studies:

- We summarize relevant works and present a methodological taxonomy regarding how they integrate structured knowledge. Including details of non-trivial operations involved in knowledge utilization.
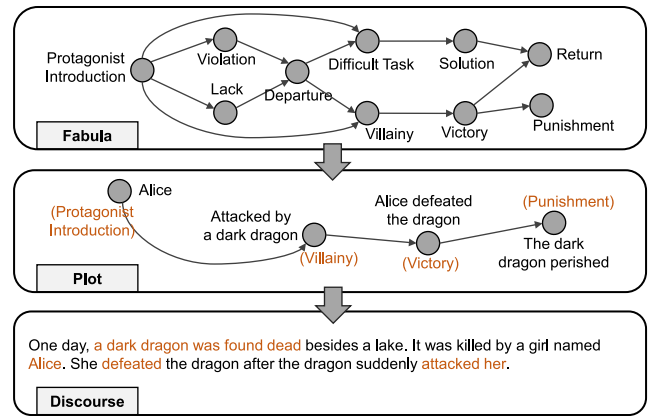

Fig. 1. Three layers of narratological concepts about story: fabula, plot, and discourse.

- We discuss available story corpora and structured knowledge datasets, as well as evaluation metrics used in story generation, along with a discussion on their characteristics from a practical point of view.
- We analyze the limitations and challenges of existing implementations, datasets, and evaluation metrics. Meanwhile, attempting to point light on future directions.

The rest of this paper is organized as follows: Section 2 presents the definitions of relevant and necessary concepts regarding stories, story generation, and structured knowledge. In Section 3, we present a taxonomy of existing works. Then, in Section 4, we discuss non-trivial aspects involved in integrating and leveraging structured knowledge in storytelling. Section 5 summarizes the available story corpora and structured knowledge datasets and presents their detailed statistics and characteristics. Section 6 presents evaluation metrics applied in story generation. In Section 7, we give insights into the current limitations and challenges from multiple aspects. Followed by a discussion on prospects for promising future research directions. Finally, Section 9 concludes this survey.

2. Terminology definition

2.1. Three conceptual layers about story

The terminology used to talk about stories in related literature, such as "narrative" and "discourse" is usually used ambiguously and hard to clarify without a specific context. Accurately understanding these terms lays the foundation for comprehending and properly discussing related works and analyzing story structure. Here we present their definitions from the narratological standpoint (adapted to use clearly distinct terms not focused only on text). There are three conceptual levels regarding stories: *fabula*, *plot*, and *discourse* [10,46]. Fig. 1 provides a top-down illustration of them.

- **Fabula**. Fabula defines the space of possible events in a story world and their logical order [47]. It can be seen as an intricate causal network [10]. Besides, it regulates story elements such as places and characters.
- **Plot**. A plot consists of a specific chain of events in a fabula and its constituent elements. It is self-contained, logically coherent, and thematically consistent. Given the same fabula, many different plots can take place.
- **Discourse**. One story can be presented to readers with various narrative manipulations or media. Discourse[2] is a particular way

---

[2] In the studies by Mieke [46] and Swartjes [10], they call the *discourse* level *narrative text* and *presentation* layer.

Table 1
Examples of four scenarios covered by story generation tasks.

Scenario: Starting sentence-prompted generation
Prompt: *Rita saw a wounded bird.* **Output:** *Her heart sank at the sight of its frailty. She cradled it tenderly and rushed to a nearest wildlife center.*

Scenario: Story ending generation
Prompt: *Tom had a puppy, but the potty training spot was untouched. A bad smell filled the room. _____* **Output:** *He found dog poop under the sofa.*

Scenario: Keyword-prompted generation
Prompt: *[cat, black bird, big tree, jump]* **Output:** *A cat spotted a black bird on a big tree. With a daring jump, it tried to catch the elusive bird, but it danced out of reach, leaving the determined cat chasing the skies.*

Scenario: Plot infilling
Prompt: *Jack woke up late this morning. _____ He feels much better now.* **Output:** *He realized he had forgotten to set his alarm. Then, he took a refreshing shower and brewed a strong cup of coffee.*

of storytelling [48]. It can completely change the effect of a story on readers. This level is what readers directly interface with.

With the rise of deep learning and large language models, which empowers a more open-world approach and less restricted scenarios, the *fabula* layer has been much less often targeted in more recent automatic story generation literature. However, it remains important in studies requiring meticulous constraints and order of plots [49–51]. Regarding the story generation task, most present works focus on generating text at the *discourse* level, while some of them touch the *plot* level, usually involving a top-down hierarchical generation process. Generally, the most common generation tasks fall into four scenarios: (1) starting sentence-prompted generation, generating subsequent text for a given story beginning; (2) story ending generation, producing a reasonable ending for a given story context; (3) keyword-prompted generation, generating a story conditioned upon a list of keywords, key-phrases, or key plots; and (4) plot infilling, generating coherent missing story pieces between given plot events. Table 1 gives examples of each of the scenarios.

2.2. Generative language model

Story generation models commonly capture and mimic word dependencies of input text to generate story continuances. A language model is the probability distribution over sequences of words, guiding how generative models predict and produce new sequences [52]. Given an input sequence $X = \{x_1, x_2, \ldots, x_n\}$, generative models aim to maximize the likelihood of generating the target sequence $Y = \{y_1, y_2, \ldots, y_m\}$. Here, each x_i and y_i is a word token. Story generation generally applies an autoregressive language model for left-to-right generation flow, of which the learning objective $\mathcal{L}_S$ is calculated as the negative likelihood of cumulative probability of words in Y:

$$\mathcal{L}_S = -\sum_{t=1}^{m} \log P(y_t | y_{<t}, X, \theta), \tag{1}$$

where $y_{<t}$ denotes the words before the tth step, θ are the trainable parameters, and m is the target sequence length.

Over the years, many Transformer-based autoregressive PLMs have been released, e.g., BART (Bidirectional and Auto-Regressive Transformers) [53], COMET (Commonsense Transformers) [54], and the GPT (Generative Pre-trained Transformer) series [55,56]; and these can be readily used as building blocks for different generation tasks. Moreover, recent non-autoregressive language models [57] can generate word sequences in parallel without modeling sequential dependencies [58]. Such models has been applied to neural machine translation and significantly speed up the generation process [59]. However, the quality

of their generated text is less satisfactory compared to autoregressive language models.

2.3. Structured knowledge

Regarding the knowledge involved in story generation, there is both *unstructured knowledge* and *structured knowledge*. Unstructured knowledge is implicit information inherent in text. Such knowledge is not shaped in a uniform form, but can be indicated and inferred from context. By contrast, structured knowledge is explicit information shaped in uniform forms. Many structured knowledge datasets have been constructed, such as triple-formed [60–63], quadruple-formed [64,65], tree-formed [66,67], and table-formed [68] datasets. Knowledge graphs are an example of triple-form datasets that collect a mass of real-world facts. Each triple is usually in the form of (*subject*, *relation*, *object*), where the subject and object entities refer to real-world objects or concepts, and the relation indicates the semantic relationship between the entities. There are many large knowledge graphs [60,61,67,69,70] storing fine-grained and instance-level knowledge like (*Albert Einstein*, *was born at*, *Ulm*). Differently, commonsense knowledge graphs [62, 63] provide general rules in the everyday world that all humans are expected to know. Such rules can be commonsense facts like (*lake*, *contains*, *water*), or causal relationships between events like (*go for a run*, *requires*, *put on running shoes*). Aside from story generation, structured knowledge has facilitated other popular tasks, like question answering [71], dialogue systems [72], recommendation systems [73], and code completion [74].

3. Integrating structured knowledge into story generation: A taxonomy

We gather and analyze existing works on structured knowledge-enhanced story generation; and classify their methods into two strategic categories regarding how they integrate structured knowledge into the generation process:

- **Structured Knowledge as Text**. Methods under this category view structured knowledge as materials that can be learned in the same way as stories. They transform structured knowledge into natural language text and feed them into generative models to capture their explicit information. This leads generative models to be inherently knowledge-enhanced.
- **Structured Knowledge as Encoding**. Methods under this category treat structured knowledge as an external instructor. They independently encode structured knowledge into low-dimensional vectors where the semantic correlation between concepts is captured. These knowledge encodings are then used to influence word probability distribution over vocabulary from which stories are generated.

We give detailed descriptions of the two strategies next. Fig. 2 presents our taxonomy, where (sub)categories and their respective works are listed.

3.1. Structured knowledge as text

For this integration strategy, the first step is to transform structured knowledge data like knowledge triples into natural language text. This transformation can be realized through simple concatenation of elements in a triple [22,42,75] or pre-defined conversion templates [37, 76]. The templates map each hard relation text into a vivid natural description, which we will discuss in Section 4. Each transformed natural language text is then divided into an input context and a predicted target. The input context consists of the subject entity and the related elements of the knowledge triple, and the predicted target contains the object entity. The input context is fed into a generative model to predict the target sequence, either by directly finetuning on PLMs, prompting,
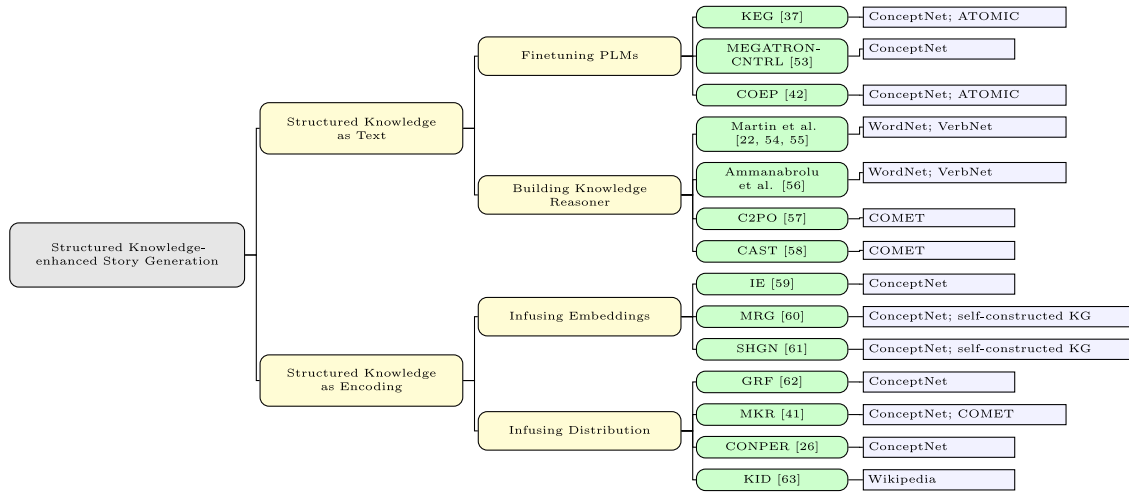
Fig. 2. A methodological taxonomy of methods on structured knowledge-enhanced story generation. Yellow boxes denote proposed (sub)categories. Green boxes denote existing works under their respective subcategories. Light blue boxes are the structured knowledge that each method utilizes.
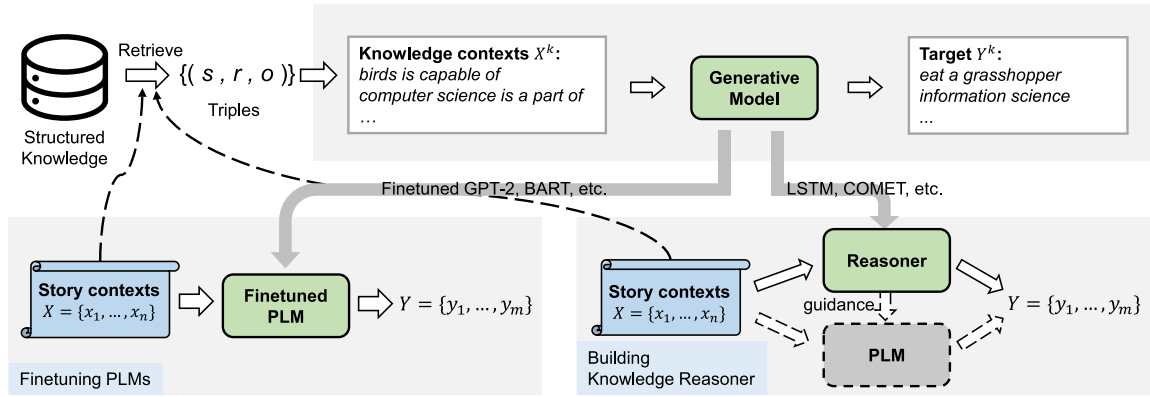

Fig. 3. Two strategies of knowledge as text category: *finetuning PLMs* and *building a knowledge reasoner*. (s, r, o) is the abbreviation for knowledge triple $(subject, relation, object)$. The dotted box and arrows indicate optional processes.

or training a knowledge generative model from scratch. We refer to the first implementations as *finetuning PLMs*, and the later implementation as *building knowledge reasoner*. Fig. 3 presents the common procedure of each. Sections 3.1.1 and 3.1.2 introduce their respective works and details.

3.1.1. Finetuning PLMs

The first work about knowledge-enhanced story generation is Knowledge-Enhanced GPT-2 (KEG) [37] proposed by Guan et al. which targets the starting sentence-prompted commonsense story generation task. It integrates the commonsense knowledge graphs ConceptNet [62] and ATOMIC [63] into GPT-2 [77], a large PLM. They post-train GPT-2 with the triple-transformed knowledge texts that are generated by templates. Formally, given a sequence of knowledge context $X^k = \{x_1^k, x_2^k, \ldots, x_q^k\}$ as input context, it is fed into GPT-2 to predict the target knowledge sequence $Y^k = \{y_1^k, y_2^k, \ldots, y_r^k\}$ through autoregressive language modeling. The learning objective $\mathcal{L}_{KG}$ is:

$$\mathcal{L}_{KG} = -\sum_{t=1}^{r} \log P(y_t^k | y_{<t}^k, X^k, \theta), \tag{2}$$

where θ are the parameters of GPT-2. The knowledge-enhanced GPT-2 is then finetuned on a story corpora by minimizing $\mathcal{L}_S$ Eq. (1). To better distinguish true and fake stories, Guan et al. additionally train a binary classifier. They shuffle, replace, and repeat sentences in true stories to create fake ones.

Following KEG, Xu et al. proposed MEGATRON-CNTRL [76] which also feeds knowledge-transformed contexts into PLMs. But instead of first finetuning PLMs on knowledge text independently, it inputs knowledge and story contexts simultaneously to generate subsequences. To be more specific, it first transforms all triples in ConceptNet into corresponding natural language sentences using templates. Next, it selects knowledge sentences related to the story. To achieve this, it extracts keywords from the input context using the RAKE algorithm [78]. A knowledge retriever then uses the extracted keywords and queries ConceptNet to retrieve related knowledge sentences. After obtaining these sentences, it further utilizes BERT [79] to encode and calculate similarity between retrieved knowledge sentences and inputted story context. The knowledge sentences with high similarity are kept and inputted into GPT-2, along with the story context, to generate the subsequent story.

COEP [42] focuses on the commonsense logic when generating a future event. Given a current story event and its previous context, it frames the story generation process into two modules: (i) an inference module that reasons knowledge explanations out of input story contexts, such as intentions and causes, and (ii) a generation module that generates future events prompted by the knowledge explanations. In its inference module, COEP finetunes BART, a PLM, with commonsense knowledge from ATOMIC in the same way as KEG. But different from transforming triples into text using pre-defined templates, COEP simply concatenates the subject entity and the relation of a triple to form the input knowledge context, then feeds it into BART to predict the

target entity. The ATOMIC-enhanced BART can generate commonsense explanations for inputted story events. Afterward, in the generation module, COEP finetunes another BART with knowledge from ConceptNet to better learn sequential dependency of events. The input story events packed with commonsense explanations are fed into the Concept-enhanced BART to predict future events.

3.1.2. Building knowledge reasoners

Instead of directly generating discourse-level stories in an end-to-end manner by finetuning PLMs, methods that build knowledge reasoners dissect the generation process. They train an independent knowledge reasoner from scratch using merely structured knowledge, which works as an event space. Thus, these knowledge reasoners are able to infer on the plot level. Then, the inferred plot lines participate in guiding discourse-level story generation.

Martin et al. [22,80,81] proposed a pipeline where story generation can roughly be broken into two steps: *event-to-event* and *event-to-sentence*. The event-to-event step generates successive events and the event-to-sequence step grounds an event to a discourse sentence. Both steps are trained with seq2seq models (e.g. RNN and LSTM). Due to the sparsity of extracted fine-grained events from input texts, Martin et al. define each event as a 4-tuple representation $\langle s, v, o, m \rangle$ where the nouns (subject s and object o) and verb (verb v) are generalized using structured knowledge datasets WordNet [67] and VerbNet [66]. m is a modifier of an event. These tuples are inputted into a seq2seq model to predict successor events. Then, another seq2seq model takes the predicted event as input and generates a sentence back to natural language discourse. Following the pipeline, Ammanabrolu et al. [75] explore an ensemble-based model in the event-to-sentence step to make generated sentences more on-topic and interesting.

Another knowledge reasoning model is COMET [54], which is built for commonsense knowledge completion. COMET is a Transformer-based model trained with knowledge triples of ConceptNet and ATOMIC. Given a seen pair consisting of a subject entity and a relation, such as (*PersonX goes to the store*, *xNeed*), COMET can complete the commonsense triple by inferring the object entity "*bring a wallet*". C2PO [82] utilizes COMET to simulate the inferential capability of humans. It targets the plot-infilling task where plot lines of a story are extracted and then need to be elaborated upon. To tackle this, C2PO employs COMET to generate intermediate events between any two consecutive plot points. These events contribute to a plot graph that contains lots of possible event flows. Then, C2PO walks a path from the starting event in the plot graph to the ending event. It calculates the probability of event co-occurrence and chooses the most possible subsequent event. After obtaining the elaborate events under the story plot lines, C2PO simply concatenates the events to form the final detailed story.

CAST [83] 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. Specifically, given a story sentence as an input context, on the one hand, CAST employs GPT-2 to generate successive sequences. On the other hand, CAST inputs the story sentence into COMET and obtains a commonsense inference explaining it, such as the character's intentions and desires. The generated sequences must conform to the inferential explanations by COMET, otherwise re-generation is required. In the confirmation stage, CAST encodes the generated sequences and the inferential explanations with sentence-BERT [84] and calculates the cosine similarity between them to measure their likeness. A whole character-centered story is generated by conducting this process iteratively.

3.2. Structured knowledge as encoding

For methods that view knowledge as encoding, they intend to capture semantic information within original knowledge graphs. Compared to converting knowledge triples into natural language sentences that can only reflect triple-level information, utilizing graph-structured knowledge can capture numerous triples within multiple hops at once. Generally, these methods first retrieve knowledge subgraphs from a knowledge base and encode subgraphs into low-dimensional embeddings using graph encoders such as graph neural networks (GNNs). Then, these embeddings are processed to twist the probability distribution over a vocabulary, so that semantic-relevant words from the knowledge graphs are more likely to be chosen.

GNNs are deep learning models for performing inference on graph data. It consists of multiple stacked aggregation layers which iteratively update the embeddings of nodes (and edges). Formally, given a knowledge graph $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ ($\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$) denotes the set of entity, relation, and triple respectively, GNNs update the embedding $\mathbf{h}_{e_i}^l$ of entity $e_i \in \mathcal{E}$ in lth layer by aggregating its neighboring and self information:

$$\mathbf{h}_{e_i}^l = \sigma \left(\frac{1}{|\mathcal{N}(e_i)|} \sum_{(e_j, r) \in \mathcal{N}(i)} \mathbf{W}_N^l \phi(\mathbf{h}_{e_j}^{l-1}, \mathbf{h}_r^{l-1}) + \mathbf{W}_S^l \mathbf{h}_{e_i}^{l-1} \right). \tag{3}$$

Here, $\mathcal{N}(e_i)$ denotes the set of directly neighboring entity-relation pairs of e_i. $\mathbf{h}_{e_i}$, $\mathbf{h}_{e_j}$, and $\mathbf{h}_r$ are latent embeddings. $\mathbf{W}_N$ and $\mathbf{W}_S$ are two learnable weight matrices. $\phi(\cdot)$ is a transition function that reflects the information passed from specific neighbors. And $\sigma(\cdot)$ is an activation function. Iterative information propagation on graphs can capture semantic relationships among entities multi-hop away.

Regarding how the output embeddings of the last layer of graph encoders can be used to twist the final word distribution and influence sequence generation, we further divide it into two implementations: *infusing embeddings* and *infusing distributions*. For the former, the embeddings are injected into manipulable generative models such as LSTMs as internal parameters. The subgraph embedding-infused generative model then directly generates subsequent sequences through the language model. The whole process happens as a black box. Differently, the implementation of infusing distributions calculates a score for each entity in the subgraph based on these embeddings. The scores measure the relevance between subgraph entities and the input story context. Later, the scores are infused with the probability distribution generated by PLMs over vocabulary as the final distribution. An overall illustration of knowledge as encoding category is presented in Fig. 4.

3.2.1. Infusing embeddings

Methods following this implementation integrate structured knowledge at the embedding level. Since manipulating the internal parameters of PLMs can cause disastrous degradation in performance [30], this implementation does not adopt PLMs as their generative models but train seq2seq models like LSTMs and Transformers from scratch. IE [85] focuses on the story ending generation task, which aims at generating an ending sentence for a story body. Given the input context, a one-hop knowledge subgraph for each conceptual word in the context is retrieved from ConceptNet. Each subgraph is then encoded by a GNN. The output embeddings of conceptual words are generated in the last layer. These embeddings later are fed into an LSTM network together with conceptual words' contextual embeddings calculated from their story context. The LSTM network tries to predict the target sentence.

MRG [86] focuses on long text generation which necessitates content planning beforehand. It constructs a domain-specific knowledge graph from input materials using Point-wise Information [87]. In the graph, each entity corresponds to a word appearing in the input texts. Entities are connected if their corresponding words appear in the same context. Also, there are weights on the edges which are calculated by
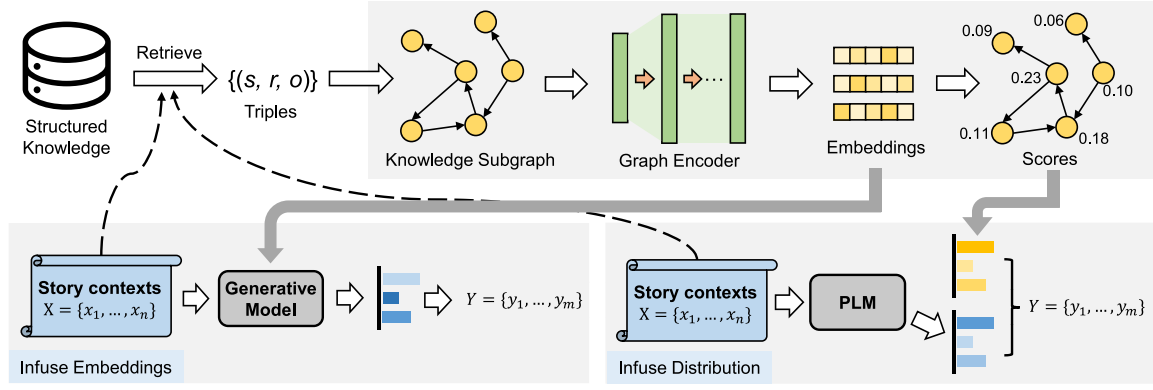
Fig. 4. Two strategies of the knowledge as encoding category: *infusing embeddings* and *infusing encodings*. (s, r, o) is the abbreviation for knowledge triple $(subject, relation, object)$. The dotted box and arrows indicate optional processes.

Table 2
Summary table of the features and techniques used in structured knowledge-enhanced story generation works. SSPG, SEG, and PI represent starting sentence-prompted generation, story ending generation, and plot infilling, respectively. The Generative Model and Generation Level columns of KID are left blank as KID is an algorithm that can be applied to various models at different generation levels.

Year	Work	Knowledge integration strategy	Task scenario	Task focus	Generative model	Generation level
2017	Martin et al. [22,80,81]	Knowledge as text	SSPG	Story event generalization	LSTM	Plot, discourse
2019	Ammanabrolu et al. [75]	Knowledge as text	SSPG	Story event grounding	LSTM	Plot, discourse
2019	IE [85]	Knowledge as encoding	SEG	Commonsense reasoning	LSTM	Discourse
2020	KEG [37]	Knowledge as text	SSPG	Knowledge awareness	GPT-2	Discourse
2020	MEGATRON-CNTRL [76]	Knowledge as text	SSPG	Story controllability	GPT-2	Discourse
2020	MRG [86]	Knowledge as encoding	SSPG	Long story generation	Transformer	Plot, discourse
2020	GRF [91]	Knowledge as encoding	SEG	Commonsense reasoning	GPT-2	Discourse
2021	C2PO [82]	Knowledge as text	PI	Commonsense reasoning	COMET	Plot, discourse
2021	CAST [83]	Knowledge as text	SSPG	Character-centric narrative	COMET	Plot, discourse
2022	COEP [42]	Knowledge as text	SSPG	Commonsense reasoning	BART	Plot, discourse
2022	CONPER [26]	Knowledge as encoding	SSPG	Persona awareness	GPT-2	Discourse
2022	SHGN [88]	Knowledge as encoding	SEG	Commonsense reasoning	Transformer	Discourse
2022	MKR [41]	Knowledge as encoding	SEG	Commonsense reasoning	COMET, BART	Discourse
2022	KID [92]	Knowledge as encoding	SEG	Knowledge awareness	–	–

word co-occurrence. These weights indicate the relevant possibility two words can co-occur. Based on the self-constructed knowledge graph, MRG conducts multi-hop reasoning and seeks possible paths as planned content. Starting from an entity, it walks to the next neighboring entity in the graph until meets an ending entity. Specifically, for each visited entity, MRG predicts whether to discard, continue, or finish this walk using a BERT-based labeler. After this, each path is represented as the sum of embeddings of all entities in the path. The path embedding worked as planned content is then fed into a Transformer together with story context to generate subsequent sequences.

Wang et al. [88] also target the story ending generation task. They propose a model named SHGN (Story Heterogeneous Graph Network) to capture both the information within the story context and the information retrieved from structured knowledge. The constructed heterogeneous graph contains three types of entities: context keywords, context sentences, and one-hop neighboring concepts of context keywords retrieved from ConceptNet. They utilize SimCSE [89], a large-scale pre-trained embedding model, to generate initial embeddings for graph entities. Then, they input the initialized embeddings into a heterogeneous graph Transformer [90] to capture the semantic relationship among entities. The final output embeddings of the heterogeneous graph Transformer are passed through Transformer decoders to generate story endings.

3.2.2. Infusing distribution

Rather than directly infusing subgraph embeddings into generative models, this strategy infuses structured knowledge at the final level of story generation, where sequences of words are selected from a probability distribution. Each entity in the knowledge subgraph is

calculated with a relevance score using embeddings. The scores reflect the relevant probability for the corresponding word over the subgraph being selected at the last stage. By combining the distribution generated by PLMs and probability distribution over words in the subgraph, this strategy twists the selection process towards knowledge words highly relevant to the input story context.

GRF [91] conducts multi-hop reasoning on commonsense knowledge subgraphs. Given an input story context, GRF recognizes concepts from the input and retrieves their neighboring concepts from ConceptNet. Then, GRF utilizes a GNN to encode entities and relations in the subgraph into low-dimensional embeddings. These embeddings participate in propagation to calculate a probability score for each concept. Specifically, GRF first assigns a score of 1 to each concept appearing in the input. Then, for other concepts, GRF updates their scores hop by hop using an embedding-based scoring function. In the last, all calculated scores are re-scaled to $(0, 1)$ by *softmax* as the local probability over concepts in the subgraph. This probability distribution is then combined with the probability distribution generated by GPT-2 over vocabulary using a gate mechanism. The subsequent sequences are then generated from the combined word distribution.

Based on GRF, MKR [41] argues that the word-level knowledge graph like ConceptNet sometimes cannot capture the semantics of events precisely. It constructs a multi-level knowledge graph, including an event-level subgraph and a word-level subgraph. To be more specific, it first extracts events from the input story context and uses COMET to generate subsequent events. All these events contribute to an event-level subgraph. Then, a word-level subgraph is constructed by dissecting each event into dispersed words which are connected by ConceptNet relations. After this, MKR encodes the multi-level knowledge

graph using a GNN. The learned embeddings are used for calculating probability scores for words in the word-level subgraph in the same way as GRF. Finally, the probability distribution over the subgraph is combined with the distribution generated by BART to generate sequential sequences.

CONPER [26] targets the story generation scenario where the generated texts should reflect the persona of the main characters. Given a story context and a piece of personality description of the main characters in the story, CONPER first extracts keywords from the story context and personality description that have an emotional tendency to accord with main characters' persona. Then, it extends these keywords by retrieving neighboring words from ConceptNet and constructs a local knowledge subgraph. Likewise, words in the knowledge subgraph are encoded into embeddings by a GNN and later fed into a full convolutional layer with *softmax* in the end to calculate the probability distribution.

Very recently, Liu et al. propose a knowledge infused decoding (KID) [92] algorithm for generative PLMs. They dynamically infuse external knowledge, which are Wikipedia word chunks, into each step of the PLM decoding. Specifically, for each story context, this work first retrieves relevant knowledge entities from a dynamically maintained knowledge memory. Then, in each step, it calculates a knowledge gain from the probabilities over the retrieved knowledge and uses the knowledge gain to define a reward that the PLM is finetuned to maximize. In a nutshell, KID reshapes the probability distribution at each decoding step toward the entities in knowledge.

3.3. Summing up

Here, we recapitulate the key features of the discussed works, encompassing their knowledge integration strategies, task scenarios, task focuses, generative models used, and the generation levels of stories. We present the summary in Table 2 and arrange these works chronologically to provide a better view of the evolution in the field.

It is easy to note that, while many of these works share a common strategy in integrating structured knowledge, they display notable diversity in terms of task scenarios and focuses. For those targeting the starting sentence-prompted generation scenario, the earliest works focus on generalizing and grounding story events by leveraging structured knowledge to produce fluent and coherent stories. However, later works shift their attention towards specific attributes of the generated text, such as knowledge awareness, the ability to generate longer narratives, and enhancing controllability of content. For works targeting story ending generation and plot infilling, they predominantly emphasize incorporating commonsense reasoning when generating text, as both of these scenarios rely heavily on causal relationships among events. Moreover, it is evident that, over time, large PLMs have become the favored backend choice as generative models. Experimental evidence showcases their advantage in capturing language patterns due to their training on vast and diverse corpora of text data (and sometimes non-textual data), outperforming non-pretrained language models significantly [37,76]. Furthermore, plot-level generation is essential for works that build knowledge reasoners into their integration strategy and those focusing on long story generation. Such intermediate level products offer overarching guidance and increase the controllability for long narrative text generation. Overall, the choice of backend model and design of integration strategies in structured knowledge-enhanced story generation works remains highly task-dependent.

4. Knowledge-related operations

Some methods described above require non-trivial operations such as extracting keywords and transforming triples into natural language sentences. These operations are critical for the effectiveness of utilizing structured knowledge. For methods that are under the knowledge as text strategy, they need to convert knowledge triples into readable texts

before processing them into generative models. For methods that utilize structured knowledge as encoding, they need to firstly extract knowledge entities from inputted story context, and then build a local knowledge subgraph. In the following, we introduce the implementations used in existing works and other practicable approaches.

4.1. Transforming triples to sentences

The simplest implementation is concatenating a triple's entities and relation. The input context in the training of COMET is basically doing so. Moreover, COMET adds a "*[MASK]*" token between the entity and relation to make them distinguishable. For example, given the triple (*natural language understanding, part of, artificial intelligence*) from ConceptNet, COMET converts to "*natural language understanding [MASK] part of [MASK] artificial intelligence*". But direct concatenation can introduce "out-of-bag" tokens to process due to the special schema of knowledge datasets, such as the relation "*xIntent*" in Atomic. Using templates can avoid such nuisance because templates convert unique tokens to readable natural language while expressing the same meaning. Levy et al. [93] provide a pre-defined set that maps each relational type $R(x, y)$ in knowledge data to at least one readable natural language sentence. For example, the relational type $educated\_at(x, y)$ can be transformed into the sentence "*x graduated from y*", where x and y are the subject entity and object entity. Other works [94–96] also provide triple-to-text rules or templates. Aside from templates, many neural network-based works [97–99] translate knowledge triples into natural language sentences by training a seq2seq model.

4.2. Extracting keywords, keyphrases, and events

One naive way to extract keywords is checking if a word's corresponding entity exists in a structured knowledge dataset. However, checking words one by one is time-consuming, especially when retrieving from vast datasets. Although some words appear in the knowledge dataset, they are trivial or unrelated to the topic of the input story. Including them can introduce noises to retrieved knowledge. To avoid these problems, IE [85] and GRF [91] only consider nouns and verbs as the candidate keywords and check them on ConceptNet. MEGATRON-CNTRL [76] utilizes RAKE algorithm [78] to extract keywords from the input context. RAKE is an unsupervised and domain-independent algorithm for extracting keywords from text. But it can still cause information loss for sometimes it cannot extract all keywords successfully [76]. CONPER [26] extracts emotional keywords from the persona description of protagonists by utilizing the NLTK sentiment analyzer. If the emotional tendency of a word is strong "*positive*" or "*negative*", this word is considered as a keyword. As for extracting events, C2PO [82] and CAST [83] use the tool OpenIE [100] to extract relational triples as events and use a pre-trained neural coreference resolution model [101] to find all character coreference. With the extracted events and character coreference, the plot lines can be built upon. Martin et al. [80] use Stanford's CoreNLP [102] to extract word dependencies that form fine-grained events. Then they generalize the fine-grained events by finding hypernyms in WordNet and VerbNet. In addition to these approaches, there are many statistical and neural network-based extraction methods for keywords [103–106], keyphrases [107–110], and events [111–114]. Since statistical approaches usually require no manually annotated training data, they are convenient to use for diverse domains [115].

4.3. Building knowledge subgraphs

There are different ways to build local knowledge subgraphs based on extracted keywords. IE [85] and CONPER [26] simply retrieve one-hop neighbors of keywords from ConceptNet to form local graphs. MRG [86] needs to find paths for content planning. Thus, it explores multi-hop neighbors in ConceptNet or self-constructed graphs with a

Table 3

Detailed statistics and characteristics of collected story corpora for story generation. Symbol "#" stands for number. "Drive" and "Causality" stands for the drive of story and causality between events, respectively.

Story Corpora	Source	# Story	# Sentence per story	# Word per story	Characteristics		
					Genre	Drive	Causality
CHILDREN'S BOOK TEST [116] https://research.facebook.com/downloads/babi	Ebooks[a]	687,343	21	464.7	Fairy tale	Character-driven	
VISUAL STORYTELLING [117] http://visionandlanguage.net/VIST	Flickr[b]	50,200	5	56.7	Realistic fiction	Plot-driven	✓
ROCSTORIES [118] https://cs.rochester.edu/nlp/rocstories	AMT writers[c]	98,161	5	49.8	Realistic fiction	Plot-driven	✓
WRITING PROMPTS [119] https://www.kaggle.com/datasets/ratthachat/writing-prompts	Reddit[d]	303,358	45.7	738.2	Fiction	Plot-driven	
FAIRY TALES [120] https://github.com/rajammanabrolu/WorldGeneration	Wikipedia	850	21.5	543.4	Fairy tale	Character-driven	
MYSTERY [120] https://github.com/rajammanabrolu/WorldGeneration	Wikipedia	532	17.6	479.4	Mystery	Plot-driven	
STORIUM [121] https://storium.cs.umass.edu	Storium[e]	5,743	1, 145.4	19, 141.5	Fiction	Plot-driven	
HIPPOCORPUS [122] http://aka.ms/hippocorpus	AMT writers[c]	6,854	17.6	292.6	Realistic fiction	Character-driven	✓
TVRECAP [123] https://github.com/mingdachen/TVRecap	Fansites[f]	29,013	9.0	147.8	Fiction	Plot-driven	
LiSCU [124] https://github.com/fabrahman/char-centric-story	Study guides[g]	1,708	48.8	1, 022.3	Fiction	Character-driven	
STORAL-EN-MO2ST [125] https://github.com/thu-coai/MoralStory	Websites[h]	1,779	17.7	302.3	Fiction	Plot-driven	✓
STORAL-ZH-MO2ST [125] https://github.com/thu-coai/MoralStory	Websites[h]	4,209	17.6	321.8	Fiction	Plot-driven	✓

[a] Ebooks from Project Gutenberg (https://www.gutenberg.org).
[b] Flickr (https://www.flickr.com) is an image and video hosting service.
[c] Human writers from Amazon Mechanical Turk (https://www.mturk.com).
[d] The r/WritingPrompts community (https://www.reddit.com/r/WritingPrompts/) on Reddit.
[e] Storium (https://storium.com) is an online creative writing game.
[f] Fan-contributed websites including Fandom (https://www.fandom.com/) and TVMegaSite (http://tvmegasite.net/).
[g] Study websites such as Shmoop (https://www.shmoop.com/) which contain educational literature material.
[h] A list of websites such as Aesop (https://www.read.gov/aesop/) which contain moral stories.

maximum hop constraint to avoid over-exploration. GRF [91] also explores multiple hops in ConceptNet, but only keeps salient concepts that are frequently visited to support information flow. Wang et al. [88] only retain concepts that are connected with keywords in more than one context sentence to control the quality of retrieved concepts. MKR [41] utilizes COMET to infer multiple subsequent events for extracted key events. The key events and inferred events build an event-level local graph. Then, MKR dissects events into words. These words together with their relational connections constitute ConceptNet form a word-level local graph.

5. Datasets

In this section, we collect available story corpora and structured knowledge datasets for story generation. We present their statistics and characters from a practical view.

5.1. Story corpora

Many story corpora covering different genres and characteristics have been constructed so far. CHILDREN'S BOOK TEST dataset [116] is originally built to measure how well language models can understand the wider linguistic context. It collects stories from 98 online children's books written by different writers. In the last sentence of each story, one word is removed for the language models to answer multiple choices. VISUAL STORYTELLING dataset [117] is the first dataset for the sequential vision-to-language task. It contains sequences of images where each image is described with an event and sequential images

constitute a coherent story. ROCSTORIES dataset [118] is a collection of five-sentence commonsense stories written by human writers. It is a widely-used benchmark for story generation for the conciseness and generality of the stories. Also, each story conveys clear causal relation between its constituted events. WRITING PROMPTS dataset [119] collects coherent and fluent passages of human-written online stories. These stories are prompted by given topics or cues. FAIRY TALES and MYSTERY datasets [120] contain the main plots of fairy tales and mystery short stories. They are extracted from the Wikipedia page. STORIUM dataset [121] collects stories from a collaborative online storytelling community. Its stories are lengthy and fine-grained with structural annotations. HIPPOCORPUS dataset [122] contains diary-like short stories of recalled and imagined events. These stories are also written by human writers. TVRECAP dataset [123] collects detailed TV show episode recaps from fan-contributed websites. Its stories feature complex interactions among multiple characters. LiSCU dataset [124] is created from various online study guides. It contains literary pieces and their summaries paired with descriptions of characters that appear in them. STORAL [125] is a collection of moral stories each of which is paired with a moral sentence. These moral stories are extracted from multiple online websites. It provides both Chinese and English moral stories.

We present the sources, statistics, and characteristics of these story datasets in Table 3. We download and calculate their total story number, average sentence number per story, and average word number per story. For the latter two figures, we calculate using sent_tokenize and word_tokenize of the nltk package. Specifically, for FAIRY TALES and MYSTERY datasets, since their original data cannot be found in their

Table 4
Detailed statistics and genres of triple-formed knowledge datasets utilized in story generation. Symbol '#' stands for the number. "Event" denotes the entities in role of base events in ATOMIC and ATOMIC2020.

Knowledge dataset	Source	# Triple	# Entity (Event)	# Relation	Genre	Instance
CONCEPTNET (v.5) [62]	OMCS	34,074,917	28,370,083	50	General	(*cat, is capable of, hunt mice*), (*book, has, knowledge*), (*water, is used for, drink*)
ATOMIC [63]	Stories, Books, Google Ngrams, Wiktionary Idioms	877,108	309,515 (24,313)	9	Inferential	(*PersonX adopts a dog, xEffect, PersonX then smiles*), (*PersonX smiles at PersonY, oReact, PersonY feel happy*)
ATOMIC-2020 [126]	Crowd-sourcing, ATOMIC, CONCEPTNET	1,331,113	638,127 (43,958)	23	inferential	(*X gets X's car repaired, happens after, X drives an old car*), (*money, has property, earned by working*)
GLUCOSE [127]	ROCSTORIES	304,099	306,903	10	Inferential	(*Times passes, causes, Someone_A grows*), (*It is winter, enables, It is cold*)
TRANSOMCS [128]	ASER	18,481,607	100,659	20	Linguistic	(*student, AtLocation, school*), (*hilarious, InstanceOf, mess*)
TUPLEKB [129]	Elementary science	282,594	6,465	1,605	Scientific	(*moose, eat, plant*), (*engine, use, energy*)

GitHub repository, we re-scrap the latest stories from the Wikipedia page using their codes and calculate their statistics. For STORAL dataset, we present the statistics of its English and Chinese stories constructed for the moral-to-story generation task, which is about generating a story given a moral sentence as a prompt. The last three columns are the characteristics of these datasets. We summarize the characteristics from three practical dimensions: (i) genre of story; (ii) drive of story; (iii) causality between events. We consider these dimensions important for choosing training corpora for a specific task. We label each story corpus with a genre like "fairy tale" or "realistic fiction". The "realistic fiction" means the stories could have actually occurred in real life. Meanwhile, a story can be driven by characters or plots. Character-driven stories mainly focus on the characters themselves, including their thoughts and conversations. While plot-driven stories focus on events, including action, cause, and effect. Furthermore, if the stories generally express strong causality between events, this corpus is marked with "✓". We also provide their download addresses.

5.2. Structured knowledge datasets

The external structured knowledge used for enhancing story generation is mostly triple-formed commonsense knowledge datasets such as CONCEPTNET and ATOMIC. The reasons why they are preferred are that, on one hand, compared to fine-grained knowledge, commonsense knowledge introduces fewer off-topic noises for a given story context. On the other hand, commonsense knowledge like ATOMIC explicitly indicates the causality between events which increases the logical coherence of stories. CONCEPTNET originated from the crowd-sourcing project OMCS (Open Mind Common Sense), with more knowledge sources such as DBpedia [60] and Wiktionary being merged over time. Each entity in CONCEPTNET denotes a commonsense concept such as "*water*" and "*drink*", and each relation indicates the semantic relation between two concepts, such as "*is used for*". The concepts are mostly words of different parts of speech, while some others are phrases. ATOMIC is an atlas containing inferential knowledge of generic events. Each entity in ATOMIC is a textual description of an event or a conclusion following an event which indicates the character's mental state or persona. The relations in ATOMIC consist of 9 kinds of "*if-event-then*" inferential relations between two entities. ATOMIC-2020 [126] enlarges ATOMIC by incorporating social and physical aspects of everyday inferential knowledge, as well as absorbing more inferential knowledge of events. The relation set of ATOMIC-2020 expands to 23 different commonsense relations, including 9 relations of social interaction, 7 physical-entity relations, and

7 event-centered relations. GLUCOSE [127] extracts 10 dimensions of causal explanation from sentences in ROCSTORIES, and generalizes them as semi-structured inference rules. Compared to ATOMIC which contains non-contextual knowledge, GLUCOSE grounds inferential knowledge to a particular story context. Furthermore, GLUCOSE extends ATOMIC's person-centric knowledge to wider topics such as places and things. TRANSOMCS [128] is a knowledge graph that collects high-quality commonsense knowledge from the linguistic graph ASER [130], whose nodes are words and edges are linguistic relations. The TupleKB [129] dataset contains high-precision, domain-specific knowledge tuples extracted from scientific text using an extraction pipeline, and guided by domain vocabulary constraints. In Table 4 we present the statistics of these triple-formed commonsense knowledge graphs and their respective commonsense genre.

Aside from commonsense knowledge graphs in triple-form, tree-formed datasets such as WordNet [67] and VerbNet [66] are used in works [22,75,80,81] to generalize events. WordNet is a large lexical database of English. It gathers synonyms of words and creates super-subordinate relations between most synonym sets such as "ISA" and "hyponymy". Synonym sets of adjectives are interlinked in terms of antonymy. Totally WordNet created around 117,000 synonym sets. VerbNet [66] is also a hierarchical network of English words but focuses only on verbs. It maps WordNet, PropBank [131], and FrameNet [132] verb types to their corresponding verb classes. Both WordNet and VerbNet can easily be read using NLTK interfaces.

6. Evaluation metrics

In this section, we summarize involved evaluation metrics in story generation and discuss their specialties. We classify these evaluation metrics into three categories: *statistical*, *neural network-based*, and *manual*. The summaries of metrics under each category are listed in Table 5.

6.1. Statistical evaluation metrics

Perplexity: The perplexity (PPL) [133] metric measures how well a language model predicts a target sequence. For a target sequence $Y = \{y_1, y_2, \ldots, y_m\}$, the PPL score is calculated as:

$$PPL(Y) = \frac{1}{\sqrt[m]{p(y_1)p(y_2)\cdots p(y_m)}}, \tag{4}$$

$$p(y_i) = \frac{\text{count}(y_i)}{\sum_{y_j \in Y} \text{count}(y_j)}. \tag{5}$$

Table 5
The list of evaluation categories and specific metrics for story generation. Arrow direction in performance indicates whether higher or lower the metric score is considered better.

Metric	Category	Performance	Description
Perplexity [133]	Statistical	↓	Inverse probability of the target sentences.
Repetition-*n* [134]	Statistical	↓	Measure the redundancy of stories. The ratio of repeated *n*-grams to all *n*-grams.
Distinct-*n* [135]	Statistical	↑	Measure the diversity of stories. The ratio of distinct *n*-grams to all *n*-grams.
Coverage [37]	Statistical	↑	The amount of commonsense triples that are covered in generated stories.
BLEU-*n* [136]	Statistical	↑	Evaluate *n*-gram overlap between two pieces of texts.
ROUGE [137]	Statistical	↑	A metric package used for evaluating the similarity between two pieces of texts.
METEOR [138]	Statistical	↑	Measure the similarity between two pieces of texts. It has features like stemming and synonymy matching.
CIDEr [139]	Statistical	↑	A metric measuring image descriptions using human consensus.
BertScore [140]	NN-based	↑	Evaluate semantic similarity between two pieces of texts which are encoded by Bert.
UNION [141]	NN-based	↑	An unreferenced metric for evaluating open-world story generation.
MAUVE [142]	NN-based	↑	Measure the gap between two pieces of texts using divergence frontiers.
Grammar [37]	Manual	↑	Indicate whether a story reads natural and fluent.
Logicality [37]	Manual	↑	Indicate whether a story is coherent to the given beginning and reasonable in terms of causal and temporal dependencies in the context.
Consistency [76]	Manual	↑	Indicate whether a story follows the same topic from beginning to end.
Informativeness [41]	Manual	↑	Assess whether the generated text produces unique and non-genetic information that is specific to the input context.
Interestingness [83]	Manual	↑	Indicate whether the story is enjoyable or not.

Here, $\text{count}(\cdot)$ is the frequency of occurrence of a word. PPL can be interpreted as the inverse probability of a target sequence normalized by the number of words. A smaller PPL score indicates better prediction performance. But for story generation, a model's confidence is hardly equal to this metric due to the diversity of gold discourse and stories. Besides, this metric is not task-specific and cannot measure the model's confidence on the event level.

Repetition-*n* and Distinct-*n*: The redundancy and diversity of generated stories can be measured with Repetition-*n* [134] and Distinct-*n* [135], with *n* is usually set to 4. Repetition-*n* calculates the percentage of tokens that repeat at *n*-gram, while distinct-*n* measures the ratio of distinct *n*-grams in generated tokens. The lower repetition-*n* and higher distinct-*n* indicate better diversity of stories. However, these two metrics also focus on word-level repetition/distinction and are unable to notice event-level and plot-level similarity/difference.

Coverage: The coverage metric [37] measures the model on how well it utilizes external knowledge. It calculates the average number of knowledge triples that appeared in each generated story. A triple is considered utilized only if its head and tail entities both appear in the generated story.

BLEU-*n*: BLEU-*n* (BiLingual Evaluation Understudy) [136] evaluates *n*-gram overlap between a generated story and a target story. Let $\hat{Y}$ and $\hat{l}$ denote the generated story sequence and its token length, and Y and l denote the target sequence and its token length. BLEU-*n* is computed as follows:

$$\text{BLEU} = BP \times \exp\left(\sum_{i=1}^{n} w_i \log p_i\left(\hat{Y}, Y\right)\right), \tag{6}$$

$$BP = \begin{cases} 1, & \text{if } l_c > l_r \text{ ;} \\ e^{1 - l_r / l_c}, & \text{if } l_c \leq l_r \text{ .} \end{cases} \tag{7}$$

Here, $p_i\left(\hat{Y}, Y\right)$, which is valued in $[0, 1]$, is the precision measuring the normalized number of *i*-gram in $\hat{Y}$ appearing in Y. w_i is the weight on *i*-gram precision. In most cases, all weights are set to $1/n$. BP is a penalty to punish generated sequences that are too short.

In story generation, *n* is usually set to 1 or 2. Caccia et al. [143] find that BLEU-*n* failed to catch the semantic or global coherence in long generated text, which results in stories with poor information content but with correct grammar that can also get a perfect BLEU score.

ROUGE: Compared with BLEU-*n* which calculates overlap between two pieces of sequences, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [137] pays more attention to how much information of the target sequence can be recalled in the generated sequence. It calculates the normalized number of *n*-grams in the target sequence that appears in the generated sequence, where *n* is set to 1 or 2.

Story generation has the notorious one-to-many issue, which means there can be plenty of good sequential sequences for the same story context but only one target sequence is given. Thus, BLEU-*n* and ROUGE are both limited for open-world story generation [119].

METEOR: The METEOR [138] improves BLEU-*n* by using WordNet for stemming and synonym matching *n*-grams. Besides, METEOR computes both precision and information recall. However, due to its utilization of WordNet, METEOR cannot evaluate generated sequences containing words outside WordNet's lexicon. And it is also inadequate for story generation, for there can be plenty of good sequential sequences aside from the target one.

CIDEr: CIDEr (Consensus-based Image Description Evaluation) [139] is originally proposed to evaluate image captions. Compared with the above *n*-gram matching evaluation metrics, CIDEr [139] further uses tf-idf (Term Frequency Inverse Document Frequency) to weight different *n*-gram when calculating similarity. Therefore, CIDEr is able to focus on "important" words in sequences.

6.2. Neural network-based evaluation metrics

With the emergence of deep neural networks, some neural network-based evaluation methods are proposed to lessen the gap between automatic evaluation with human judgments.

BERTScore: Unlike previous methods that calculate token-level syntactic similarity, BERTScore [140] computes the similarity between the generated sequence and the target sequence using their contextual embeddings generated by BERT. The scores are calculated with pairwise cosine similarity between two embeddings. BERTScore is much more flexible in matching two sequences' semantics. However, the grammar errors might be ignored due to the comparison between only embedding not words.

UNION: UNION [141] is a Bert-based metric that does not need a target sequence for reference to measure the generated story. It is trained as a classifier aiming to distinguish human-written stories from negative samples. These negative samples are created with manipulations including repetition, substitution, reordering, and negation alteration. Thus, UNION has the ability to capture commonly observed issues like repeated plots and conflicting logic in generated stories.

MAUVE: MAUVE [142] compares the learned distribution from a language model to the distribution of human-written texts using divergence frontiers. It can reflect the quality of generated stories affected by length, decoding algorithm, and model size. These properties are not captured and analyzed in existing metrics.

6.3. Manual evaluation metrics

Though manual evaluation is costly but still necessary for evaluating generated stories due to the limitations of automatic metrics [36]. A service for manual evaluation is Amazon Mechanical Turk[3] (AMT), where annotators are asked questions covering the following dimensions of story quality and give a score on each dimension.

Grammar: Measuring the grammar correctness of generated texts such as subject-verb agreement, tense, singular/plural agreement, punctuation, and sentence fragments.

Logicality: This measures whether the events, characters' actions, and intentions in generated stories accord with the input contexts and are reasonable. A good story should act up to audiences' inference. However, the logicality is hard to measure for automatic metrics.

Consistency: This measuring whether the generated stories and given context are on the same topic and under persistent rules. Some inconsistencies in a story are like protagonists' "out of character", the settings behind stories contradict before and after.

Informativeness: This dimension reflects how much knowledge users can perceive when reading the generated stories. The coverage metric is designed for a similar purpose by counting knowledge triples appearance without being aware of knowledge relevance. Given two pieces of generated sequences, they may contain the same number of knowledge triples, but their informativeness can vary much due to different choices of triples.

Interestingness: Whether a story is interesting is hard to evaluate but essential to readers. A story might be reasonable and fluent but uninteresting. Nowadays, studies just let human judges distinguish whether a story is interesting or not according to their tastes, which can vary from individual to individual. Maybe there is still a long way to go to measure the story's interestingness automatically.

7. Discussion on challenges

Incorporating structured knowledge has significantly advanced open-world story generation, but the quality of generated stories is still far from human-written counterparts. To bridge the gap, efforts are required to tackle multi-aspect challenges that hinder the way. The challenges include the inherent limitation of PLMs, the information loss when utilizing structured knowledge, and even the evaluation metrics. In the following, we dissect each challenge and discuss solutions for them.

7.1. Limitation of PLMs for storytelling

Although PLMs have demonstrated remarkable text generation capabilities, the generated text has shown the prevalence of being generic, including the scope and precision of word choosing [144]. This is due to the training materials of PLMs, which are crowd-sourced from broad genres while falling short of reflecting the characteristics of story narration. Overcoming this inherent limitation of PLMs is challenging, as it incurs significant costs to train or fine-tune the models [55]. In many research domains, such as bio-medicine, engineering, and finance, people began to train domain-specific PLMs on respective resources [145–147], and these models have brought great advances in each field. Likewise, we can train storytelling-specific PLMs exclusively on story corpora to make the generated narratives more story resembled. Though some existing works like MRG [86] and SHGN [88] train a Transformer decoder from scratch, they are corpus-oriented and thus produce stories with certain narration styles. However, to realize storytelling-specific PLMs, training on large-scale and diverse corpora is necessary.

7.2. Limitation of knowledge sources

Another impediment is the limited structured knowledge sources. Almost all existing works on structured knowledge-enhanced story generation only utilize commonsense knowledge datasets like CONCEPT-NET and ATOMIC. Although commonsense knowledge inherits abundant generic facts and inferential explanations, which can boost stories' logical coherence and increase their readability, making stories more informative and descriptive can be tricky. For example, imagine generating a story about a real person such as *Emily Dickinson*; if the model can be aware of Dickinson's family background, occupation, and literary works, the generated story can be more convincing and vivid. Unfortunately, commonsense knowledge datasets do not contain such detailed facts to elaborate on. Also, PLMs show unsatisfactory performance in expressing fine-grained knowledge because they are trained with online texts which overgeneralize or contain big noises for specific knowledge.

Incorporating fine-grained knowledge is challenging because such knowledge datasets, like DBpedia [60] and Wikidata [68], are enormous. Each entity usually links many varied neighboring entities and attributes. This causes retrieving information from these knowledge datasets to introduce irrelevant noise from the given story context. Besides, since fine-grained knowledge is not generic and needs to cater to specific contexts, it requires real-time retrieval during generation instead of being finished while preprocessing. Such real-time retrieval from vast databases can increase generation time costs. Overall, a trade-off is necessary when incorporating fine-grained knowledge.

7.3. Information loss in knowledge-related operations

During the intermediate generation process, non-trivial operations such as keyword extraction and triple-to-text conversion can cause information loss, hindering knowledge enhancement's desired performance boost. Many methods [37,42,76] transform knowledge triples into texts by pre-defined templates. While such templates can yield readable texts, they are limited in linguistic and creative diversity. For each relation, templates only paraphrase it to one or a few natural language phrases, which is far more monotonous than how diversely and vividly meaning can be conveyed in the real world. Because language models generate words according to the word correlation pattern in training data, template-based transformation may have difficulty sensing and elaborating upon knowledge when encountering a hint expressed outside pre-defined templates. Thus, more vivid, diverse, and robust approaches to transforming knowledge triples into readable texts are needed.

Extracting keywords or key events from a story context to retrieve relevant knowledge can also involve information loss and noise. If all words appearing in knowledge datasets are extracted, some of them will introduce off-topic information because not every word is a keyword. Guan et al. [37] and GRF [91] tackle this by only extracting nouns and verbs. Meanwhile, extracting key events can produce inaccurate results and causes a detrimental effect on performance. Therefore, more reliable extraction tools and algorithms are necessary to ensure the accuracy of extracted information.

7.4. Limitation on controllability

Since PLMs produce token sequences from learned language models, their generated stories accord with the distribution of crowd-sourced data and usually deviate from readers' expectations [44,148]. Readers may deem such unexpectedness as creative because they are unfamiliar with the generated samples. However, PLMs, in many circumstances, can be seen to mimic what they see in the training data [149–151]. As for readers' expectations, after they input a starting sentence, a list of keywords, or the plot lines of a story, in many cases, they have anticipations for story continuation, such as specific actions and

endings. How to control generative models to produce stories satisfying readers' expectations is trending in story generation studies. Enhancement by structured knowledge is one of the manners of controllability, for it encourages the generated stories to reflect relevant knowledge to the input context. It helps to keep the stories on topic and coherent. Nevertheless, reflecting on relevant knowledge is still not enough. Some studies try to make hard control on PLMs to produce guided words or plots from users' input. Dathathri et al. [30] combine a PLM with a simple attribute classifier consisting of a user-specified bag of words. This classifier alters the latent representations of input words and pushes the PLM's hidden activation to guide the generation. PlotMachine [28] transforms a plot outline into a coherent story by tracking the dynamic plot states. It loosely integrates key plots into the output narrative. Pascua et al. [152] add a shift to the probability distribution over vocabulary towards given guided words. It modifies the log probability of the words according to their semantic similarity to the guide words, quantified as the cosine similarity between their vector representations. Soft control aims at directing the mood or high-level topic of stories, which is more difficult but sensible. Some studies [25,26] try to regulate the personalities or emotions of protagonists of stories. However, more kinds of controllability are left to explore.

7.5. Limitation on commonsense inference

Commonsense inference in the text is about finding the relatedness between narrative events and their causal relationship. Considering two events "*win a lottery*" and "*feel excited*", one can easily infer that the former event is likely to "*result in*" the latter. Conveying a clear inference between story events is critical for good readability and coherence. Many works chose to finetune PLMs on inferential datasets like ATOMIC to achieve better inference awareness in generated stories. However, as Liu et al. highlight in their work [153], finetuning PLMs directly on inferential datasets can mess up both the language generation process and the commonsense inference process. They propose a new task named Context Commonsense Inference in Sentence Selection (CIS^2) to examine models' ability on commonsense inference better. This task abstracts away language generation and forces the language model to focus exclusively on inference. But still, presenting awareness of commonsense inference is an open and challenging problem in story generation and other generation tasks [154].

7.6. Limitations of automatic evaluation metrics

Another important issue for story generation is the lack of robust machine evaluation metrics. Although many automatic evaluation metrics listed in Table 5 have been used in existing works, they cannot accurately reflect the quality of generated stories. These evaluations correlate poorly with human judgments [155,156]. These metrics either score the similarity between generated stories with "golden answers" (Perplexity, BLEU-n, ROUGE, METEOR, CIDEr, BertScore, MAUVE), measure how many knowledge triples appear in generated stories (Coverage), or check verbal repetition and diversity (Repetition-n, Distinct-n). However, as is well-known, there is no golden answer to what is a good story, and perhaps there are infinite narratives to make a story good [157]. Relying solely on one evaluation metric may underestimate interesting and good-quality stories. Thus, recruiting human judges to rate stories is indispensable for fair evaluation, despite the cost. Recently, some hybrid-metric systems like ADEM [158], BLEURT [159], and STORIUM [121] aim to evaluate stories more comprehensively. Nonetheless, more reliable evaluation metrics are needed. One inspiration can be imitating humans' mental processes [160,161].

Guan et al. propose a benchmark OpenMEVA [162] for automatic story generation metrics. OpenMEVA provides a comprehensive test suite to assess the capabilities of different metrics. They find that existing metrics all have a low correlation with human judgments and difficulty noticing off-topic and conflicting plots. Xie et al. propose an approach named DeltaScore that utilizes perturbation to evaluate fine-grained story aspects [163]. Their approach demonstrates a strong correlation with human evaluations.

8. Future directions

8.1. Non-text storytelling

Open-world storytelling can be conveyed through various mediums. Recently, some compelling pre-trained generative models can create realistic images and art from a description in natural language. Examples include DALLE-1 [164], DALLE-2 [165], and Stable Diffusion [166]. Incorporating text story generation with text-to-image generative models can present creative and vivid stories. An interesting application of this approach is creating illustrations for generated stories, also known as a "storyboard". However, existing pragmatic storyboard systems are mostly based on retrieving images for story text [167–170]. They are short of creative and flexible storytelling ability. In 2023, Chen et al. propose a story generation pipeline for visual storytelling [171] that includes a narrative generation section and an image generation section. At around the same time, Hong et al. proposed a character-based visual storytelling model driven by coherence [172].

8.2. Different narrative techniques

Most open-world story generation works settle on chronological storytelling, where plots are presented in temporal order. However, evidence has suggested that different discourse organizations of underlying events can affect readers' mental processes while reading and comprehending [173]. Non-chronological narrative techniques like flashbacks and flashforwards can stir readers' curiosity. Targeting this, Han et al. [174] propose a hierarchical model that generates flashbacks in stories by encoding temporal dependencies of events as temporal prompts and inputting them into the generative model. Other narrative techniques like foreshadowing, plot twists, and multiple points of view deserve more attention for application.

8.3. Collaborative storytelling

Most story generation tasks are about completing a story given an existing context, and little human-machine collaboration occurs except for inputting prompts from time to time. However, collaborative storytelling enables humans and machines to mutually assist in their creativity [175]. Several works have made efforts to incorporate more human creativity in automatic storytelling. Goldfarb-Tarrant et al. [175] propose an interactive system that allows flexible human involvement in story planning, editing, and revising without regeneration of the whole story. Brahman et al. [176] focus on the scenario that users constantly provide cue phrases to drive the generation process. Other works such as Dramatron [177] achieve collaborative storytelling in specific domains like screenplays and theater scripts [178,179]. More sophisticated collaborations, such as recognizing discourse-level causal relations and character emotions, need further study.

8.4. Generating long stories

Existing story generation practices mostly target very short stories, which can be noticed from Table 3. The stories used in their works are often less than 1000 words. While in the real world, the stories humans read are actually much longer, and the limitation on length is a major obstacle for automatic story generation to be in a more extensive application. Generating long stories is difficult. The story needs to maintain a coherent throughout plot, preserve the same narrative style, and avoid inconsistency. Recently, Yang et al. [180] propose a Recursive Reprompting and Revision framework to generate long stories. As the first work to explicitly target long story generation, this work leaves promising room for further improvement and exploration, including a more abundant outline plan and more comprehensive inconsistency detection.

9. Conclusion

In this survey, we discuss the progress in the field of structured knowledge-enhanced story generation. As the main contribution of this survey, we discuss the technical details of related works and present a systematic taxonomy regarding how they incorporate structured knowledge into story generation process: *structured knowledge as text* and *structured knowledge as encoding*. Through a comparison of the features of different works, we call attention to the diverse scenarios and focuses of approaches to story generation tasks. The integration strategies and model designs of the analyzed works are highly task-dependent, making it unfeasible to put them under the same evaluation metrics. We summarize the characteristics of story corpora, knowledge datasets, and their evaluation metrics as contributions to foster further research. Likewise, we raise insights into the current challenges of open-world story generation and cast light on promising directions for future study, especially on what external knowledge sources to use and how to better leverage such knowledge. We hope this survey aid readers in understanding the latest developments in this field and inspire new research avenues.

CRediT authorship contribution statement

Yuxin Wang: Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Writing – review & editing. **Jieru Lin:** Formal analysis, Resources. **Zhiwei Yu:** Conceptualization, Methodology. **Wei Hu:** Supervision, Writing – original draft, Writing – review & editing. **Börje F. Karlsson:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] R. Schank, Tell Me a Story: Narrative and Intelligence, Northwestern University, 1995.

[2] F. Polletta, P.C.B. Chen, B.G. Gardner, A. Motes, The sociology of storytelling, Annu. Rev. Sociol. 37 (1) (2011) 109–130, http://dx.doi.org/10.1146/annurev-soc-081309-150106.

[3] H. McLellan, Corporate storytelling perspectives, J. Qual. Participation 29 (1) (2006) 17.

[4] C.E. Abrahamson, Storytelling as a pedagogical tool in higher education, Education 118 (3) (1998) 440–452.

[5] H. Jenkins, Transmedia storytelling and entertainment: An annotated syllabus, in: Entertainment Industries, Routledge, 2014, pp. 153–168, http://dx.doi.org/10.1080/10304312.2010.510599.

[6] J. Wallis, Making games that make stories, in: Second Person, MIT Press, 2007.

[7] A.I. Alhussain, A.M. Azmi, Automatic story generation: A survey of approaches, ACM Comput. Surv. 54 (5) (2021) http://dx.doi.org/10.1145/3453156.

[8] D. Wilmot, Great expectations: Unsupervised inference of suspense, surprise and salience in storytelling, 2022, http://dx.doi.org/10.48550/arXiv.2206.09708.

[9] B.F. Karlsson, A.L. Furtado, F.G. Guerra, On the Craft of Interactive Stories, Tech. Rep. MCC 36/09, Informatics Dept., PUC-Rio, 2009, URL http://www.dbd.puc-rio.br/depto_informatica/09_36_karlsson.pdf.

[10] I. Swartjes, M. Theune, A fabula model for emergent narrative, in: TIDSE, 2006, pp. 49–60, http://dx.doi.org/10.1007/11944577_5.

[11] B.F. Karlsson, A.L. Furtado, Conceptual model and system for genre-focused interactive storytelling, in: Entertainment Computing - ICEC 2014, Springer, 2014, pp. 27–35.

[12] J.R. Meehan, TALE-SPIN, an interactive program that writes stories, in: Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Vol. 77, 1977, pp. 91–98.

[13] S.R. Turner, Minstrel: A Computer Model of Creativity and Storytelling, University of California, Los Angeles, 1993.

[14] N.M. Sgouros, Dynamic generation, management and resolution of interactive plots, Artificial Intelligence 107 (1) (1999) 29–62.

[15] R.M. Young, et al., An overview of the mimesis architecture: Integrating intelligent narrative control into an existing gaming environment, in: Working Notes of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment, AAAI Press, 2001, pp. 77–81.

[16] N. Szilas, Structural models for interactive drama, in: COSIGN, 2002, p. 22.

[17] M. Mateas, A. Stern, Structuring content in the Façade interactive drama architecture, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Vol. 1, no. 1, 2005, pp. 93–98.

[18] H. Barber, D. Kudenko, Dynamic generation of dilemma-based interactive narratives, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Vol. 3, no. 1, 2007, pp. 2–7.

[19] B. Magerko, Story representation and interactive drama, in: Proc. of 1st Artificial Intelligence and Interactive Digital Entertainment Conference, Citeseer, 2005, pp. 87–92.

[20] Y.-G. Cheong, R.M. Young, Suspenser: A story generation system for suspense, IEEE Trans. Comput. Intell. AI Games 7 (1) (2014) 39–52.

[21] V.I. Propp, Morphology of the Folktale, Vol. 9, University of Texas Press, 1968.

[22] L.J. Martin, P. Ammanabrolu, X. Wang, S. Singh, B. Harrison, M. Dhuliawala, P. Tambwekar, A. Mehta, R. Arora, N. Dass, et al., Improvisational storytelling agents, in: NeurIPS Workshop, 2017, p. 4.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, Vol. 30, 2017, pp. 6000–6010, http://dx.doi.org/10.5040/9781350101272.00000005.

[24] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving Language Understanding by Generative Pre-Training, OpenAI, 2018.

[25] F. Brahman, S. Chaturvedi, Modeling protagonist emotions for emotion-aware storytelling, in: EMNLP, 2020, pp. 5277–5294, http://dx.doi.org/10.18653/v1/2020.emnlp-main.426.

[26] Z. Zhang, J. Wen, J. Guan, M. Huang, Persona-guided planning for controlling the protagonist's persona in story generation, in: NAACL, 2022, pp. 3346–3361, http://dx.doi.org/10.18653/v1/2022.naacl-main.245.

[27] Y. Mori, H. Yamane, Y. Mukuta, T. Harada, Computational storytelling and emotions: A survey, 2022, http://dx.doi.org/10.48550/arXiv.2205.10967, arXiv.

[28] H. Rashkin, A. Celikyilmaz, Y. Choi, J. Gao, PlotMachines: Outline-conditioned generation with dynamic plot state tracking, in: EMNLP, 2020, pp. 4274–4295, http://dx.doi.org/10.18653/v1/2020.emnlp-main.349.

[29] S. Goldfarb-Tarrant, T. Chakrabarty, R. Weischedel, N. Peng, Content planning for neural story generation with aristotelian rescoring, in: EMNLP, 2020, pp. 4319–4338, http://dx.doi.org/10.18653/v1/2020.emnlp-main.351.

[30] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, R. Liu, Plug and play language models: A simple approach to controlled text generation, in: ICLR, 2020, http://dx.doi.org/10.48550/arXiv.1912.02164.

[31] B. Tan, Z. Yang, M. Al-Shedivat, E. Xing, Z. Hu, Progressive generation of long text with pretrained language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4313–4324, http://dx.doi.org/10.18653/v1/2021.naacl-main.341.

[32] E. Nichols, L. Gao, R. Gomez, Collaborative storytelling with large-scale neural language models, in: MIG, 2020, pp. 1–10, http://dx.doi.org/10.48550/arXiv.2011.10208.

[33] Z. Zhang, Y. Xu, Y. Wang, B. Yao, D. Ritchie, T. Wu, M. Yu, D. Wang, T.J.-J. Li, StoryBuddy: A human-AI collaborative chatbot for parent-child interactive storytelling with flexible parental involvement, in: CHI, 2022, pp. 1–21, http://dx.doi.org/10.1145/3491102.3517479.

[34] A. Yuan, A. Coenen, E. Reif, D. Ippolito, Wordcraft: Story writing with large language models, in: IUI, 2022, pp. 841–852, http://dx.doi.org/10.1145/3490099.3511105.

[35] G. Chen, Y. Liu, H. Luan, M. Zhang, Q. Liu, M. Sun, Learning to generate explainable plots for neural story generation, TASLP 29 (2021) 585–593, http://dx.doi.org/10.1109/TASLP.2020.3039606.

[36] A. See, A. Pappu, R. Saxena, A. Yerukola, C.D. Manning, Do massively pretrained language models make better storytellers? in: CoNLL, 2019, pp. 843–861, http://dx.doi.org/10.18653/v1/K19-1079.

[37] J. Guan, F. Huang, Z. Zhao, X. Zhu, M. Huang, A knowledge-enhanced pretraining model for commonsense story generation, TACL 8 (2020) 93–108, http://dx.doi.org/10.1162/tacl_a_00302.

[38] K. Shirai, K. Hashimoto, A. Eriguchi, T. Ninomiya, S. Mori, Neural text generation with artificial negative examples to address repeating and dropping errors, JNLP 28 (3) (2021) 751–777, http://dx.doi.org/10.5715/jnlp.28.751.

[39] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, R. Yan, Knowledge-grounded dialogue generation with pre-trained language models, in: EMNLP, 2020, pp. 3377–3390, http://dx.doi.org/10.18653/v1/2020.emnlp-main.272.

[40] Z. Lin, M. Riedl, Plug-and-blend: A framework for controllable story generation with blended control codes, in: Proceedings of the Third Workshop on Narrative Understanding, 2021, pp. 62–71, http://dx.doi.org/10.18653/v1/2021.nuse-1.7.

[41] F. Mu, W. Li, Enhancing text generation via multi-level knowledge aware reasoning, in: IJCAI, 2022, pp. 4310–4316, http://dx.doi.org/10.24963/ijcai.2022/598.

[42] L. Lin, Y. Cao, L. Huang, S. Li, X. Hu, L. Wen, J. Wang, What makes the story forward? Inferring commonsense explanations as prompts for future event generation, in: SIGIR, 2022, pp. 1098–1109, http://dx.doi.org/10.1145/3477495.3532080.

[43] C. Hou, C. Zhou, K. Zhou, J. Sun, S. Xuanyuan, A survey of deep learning applied to story generation, in: SmartCom, 2019, pp. 1–10, http://dx.doi.org/10.1007/978-3-030-34139-8_1.

[44] A. Alabdulkarim, S. Li, X. Peng, Automatic story generation: Challenges and attempts, in: Third Workshop on Narrative Understanding, 2021, pp. 72–83, http://dx.doi.org/10.18653/v1/2021.nuse-1.8.

[45] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, M. Jiang, A survey of knowledge-enhanced text generation, ACM Comput. Surv. (2022) http://dx.doi.org/10.1145/3512467.

[46] M. Bal, C. Van Boheemen, Narratology: Introduction to the Theory of Narrative, University of Toronto Press, 2009.

[47] L. Castricato, S. Frazier, J. Balloch, M. Riedl, Fabula entropy indexing: Objective measures of story coherence, in: Workshop on Narrative Understanding, 2021, pp. 84–94, http://dx.doi.org/10.18653/v1/2021.nuse-1.9.

[48] H. White, The Content of the Form: Narrative Discourse and Historical Representation, JHU Press, 1990.

[49] R. Cantoni, J. Essenziale, M. Pezzera, R. Mainetti, N.A. Borghese, Procedural constrained story generation based on propp's and fabula models, in: SeGAH, 2020, pp. 1–8, http://dx.doi.org/10.1109/SeGAH49190.2020.9201843.

[50] S.G. Ware, R.M. Young, B. Harrison, D.L. Roberts, A computational model of plan-based narrative conflict at the fabula level, TCIAIG 6 (3) (2014) 271–288, http://dx.doi.org/10.1109/TCIAIG.2013.2277051.

[51] F.P. Santos, Changing an organization's legitimation story: Navigating between the materiality of the past and the strategy for the future, Culture Organ. (2022) 1–24, http://dx.doi.org/10.1080/14759551.2022.2103133.

[52] A. Mnih, G. Hinton, A scalable hierarchical distributed language model, in: NeurIPS, 2008, pp. 1081–1088, http://dx.doi.org/10.5555/2981780.2981915.

[53] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: ACL, 2020, pp. 7871–7880, http://dx.doi.org/10.18653/v1/2020.acl-main.703.

[54] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, COMET: Commonsense transformers for automatic knowledge graph construction, in: ACL, 2019, pp. 4762–4779, http://dx.doi.org/10.18653/v1/P19-1470.

[55] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, Language models are few-shot learners, in: NeurIPS, Vol. 33, 2020, pp. 1877–1901, http://dx.doi.org/10.48550/arXiv.2005.14165.

[56] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022, arXiv:2203.02155.

[57] J. Gu, J. Bradbury, C. Xiong, V.O. Li, R. Socher, Non-autoregressive neural machine translation, in: ICLR, 2018, http://dx.doi.org/10.48550/arXiv.1711.02281.

[58] K. Yang, W. Lei, D. Liu, W. Qi, J. Lv, POS-constrained parallel decoding for non-autoregressive generation, in: ACL, 2021, pp. 5990–6000, http://dx.doi.org/10.18653/v1/2021.acl-long.467.

[59] Y. Xiao, L. Wu, J. Guo, J. Li, M. Zhang, T. Qin, T.-y. Liu, A survey on non-autoregressive generation for neural machine translation and beyond, 2022, http://dx.doi.org/10.48550/arXiv.2204.09269, arXiv.

[60] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia, Semantic Web 6 (2015) 167–195, http://dx.doi.org/10.3233/SW-140134.

[61] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: A collaboratively created graph database for structuring human knowledge, in: SIGMOD, 2008, pp. 1247–1250, http://dx.doi.org/10.1145/1376616.1376746.

[62] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, in: AAAI, 2017, pp. 4444–4451, http://dx.doi.org/10.48550/arXiv.1612.03975.

[63] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N.A. Smith, Y. Choi, ATOMIC: An atlas of machine commonsense for if-then reasoning, in: AAAI, Vol. 33, 2019, pp. 3027–3035, http://dx.doi.org/10.1609/aaai.v33i01.33013027.

[64] A. García-Durán, S. Dumančić, M. Niepert, Learning sequence encoders for temporal knowledge graph completion, in: EMNLP, 2018, pp. 4816–4821, http://dx.doi.org/10.18653/v1/D18-1516.

[65] K. Leetaru, P.A. Schrodt, Gdelt: Global data on events, location, and tone, in: ISA Annual Convention, Vol. 2, 2013, pp. 1–49, http://dx.doi.org/10.5860/choice.52-0044.

[66] K.K. Schuler, VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon, University of Pennsylvania, 2005.

[67] G.A. Miller, WordNet: A lexical database for English, Commun. ACM 38 (11) (1995) 39–41, http://dx.doi.org/10.1145/219717.219748.

[68] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85, http://dx.doi.org/10.1145/2629489.

[69] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka, T.M. Mitchell, Toward an architecture for never-ending language learning, in: AAAI, 2010, pp. 1306–1313.

[70] F. Mahdisoltani, J. Biega, F. Suchanek, Yago3: A knowledge base from multilingual Wikipedias, in: CIDR, 2014.

[71] Y. Lan, G. He, J. Jiang, J. Jiang, W.X. Zhao, J.-R. Wen, A survey on complex knowledge base question answering: Methods, challenges and solutions, in: IJCAI, 2021, pp. 4483–4491, http://dx.doi.org/10.24963/ijcai.2021/611.

[72] J. Ni, T. Young, V. Pandelea, F. Xue, E. Cambria, Recent advances in deep learning based dialogue systems: A systematic survey, Artif. Intell. Rev. (2022) 1–101.

[73] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, A survey on knowledge graph-based recommender systems, TKDE 34 (8) (2022) 3549–3568, http://dx.doi.org/10.1109/TKDE.2020.3028705.

[74] S. Lu, N. Duan, H. Han, D. Guo, S.-w. Hwang, A. Svyatkovskiy, ReACC: A retrieval-augmented code completion framework, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 6227–6240, http://dx.doi.org/10.18653/v1/2022.acl-long.431.

[75] P. Ammanabrolu, E. Tien, W. Cheung, Z. Luo, W. Ma, L. Martin, M. Riedl, Guided neural language generation for automated storytelling, in: Storytelling Workshop, 2019, pp. 46–55, http://dx.doi.org/10.18653/v1/W19-3405.

[76] 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, in: EMNLP, 2020, pp. 2831–2845, http://dx.doi.org/10.18653/v1/2020.emnlp-main.226.

[77] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, OpenAI, 2019.

[78] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text Min.: Appl. Theory 1 (2010) 10–1002, http://dx.doi.org/10.1002/9780470689646.ch1.

[79] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[80] L. Martin, P. Ammanabrolu, X. Wang, W. Hancock, S. Singh, B. Harrison, M. Riedl, Event representations for automated story generation with deep neural nets, in: AAAI, Vol. 32, 2018, http://dx.doi.org/10.48550/arXiv.1706.01331.

[81] L.J. Martin, S. Sood, M.O. Riedl, Dungeons and dqns: Toward reinforcement learning agents that play tabletop roleplaying games, in: INT/WICED@ AIIDE, 2018, URL http://ceur-ws.org/Vol-2321/paper4.pdf.

[82] P. Ammanabrolu, W. Cheung, W. Broniec, M.O. Riedl, Automated storytelling via causal, commonsense plot ordering, AAAI 35 (7) (2021) 5859–5867, http://dx.doi.org/10.48550/arXiv.2009.00829.

[83] X. Peng, S. Li, S. Wiegreffe, M. Riedl, Inferring the reader: Guiding automated story generation with commonsense reasoning, 2021, http://dx.doi.org/10.48550/arXiv.2105.01311, CoRR.

[84] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: EMNLP-IJCNLP, 2019, pp. 3982–3992, http://dx.doi.org/10.18653/v1/D19-1410.

[85] J. Guan, Y. Wang, M. Huang, Story ending generation with incremental encoding and commonsense knowledge, in: AAAI, Vol. 33, 2019, pp. 6473–6480, http://dx.doi.org/10.1609/aaai.v33i01.33016473.

[86] L. Zhao, J. Xu, J. Lin, Y. Zhang, H. Yang, X. Sun, Graph-based multi-hop reasoning for long text generation, 2020, http://dx.doi.org/10.48550/arXiv.2009.13282.

[87] A. Morris, J.-L. Schwartz, P. Escudier, An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram, Comput. Speech Lang. 7 (2) (1993) 121–136, http://dx.doi.org/10.1006/csla.1993.1006.

[88] J. Wang, B. Zou, Z. Li, J. Qu, P. Zhao, A. Liu, L. Zhao, Incorporating commonsense knowledge into story ending generation via heterogeneous graph networks, 2022, http://dx.doi.org/10.48550/arXiv.2201.12538, arXiv.

[89] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: EMNLP, 2021, pp. 6894–6910, http://dx.doi.org/10.18653/v1/2021.emnlp-main.552.

[90] Z. Hu, Y. Dong, K. Wang, Y. Sun, Heterogeneous graph transformer, in: WWW, 2020, pp. 2704–2710, http://dx.doi.org/10.1145/3366423.3380027.

[91] H. Ji, P. Ke, S. Huang, F. Wei, X. Zhu, M. Huang, Language generation with multi-hop reasoning on commonsense knowledge graph, in: EMNLP, 2020, pp. 725–736, http://dx.doi.org/10.18653/v1/2020.emnlp-main.54.

[92] R. Liu, G. Zheng, S. Gupta, R. Gaonkar, C. Gao, S. Vosoughi, M. Shokouhi, A.H. Awadallah, Knowledge infused decoding, in: ICLR, 2022, http://dx.doi.org/10.48550/arXiv.2204.03084.

[93] O. Levy, M. Seo, E. Choi, L. Zettlemoyer, Zero-shot relation extraction via reading comprehension, in: CoNLL, 2017, pp. 333–342, http://dx.doi.org/10.18653/v1/K17-1034.

[94] P. Cimiano, J. Lüker, D. Nagel, C. Unger, Exploiting ontology lexica for generating natural language texts from RDF data, in: ENLG Workshop, 2013, pp. 10–19.

[95] R. Dale, S. Geldof, J.-P. Prost, CORAL: Using natural language generation for navigational assistance, in: ACSC, Australian Computer Society, Inc., 2003, pp. 35–44, http://dx.doi.org/10.5555/783106.783111.

[96] R. Turner, Y. Sripada, E. Reiter, Generating approximate geographic descriptions, in: ENLG Workshop, Association for Computational Linguistics, 2009, pp. 42–49.

[97] P. Vougiouklis, H. Elsahar, L.-A. Kaffee, C. Gravier, F. Laforest, J. Hare, E. Simperl, Neural Wikipedian: Generating textual summaries from knowledge base triples, J. Web Semant. 52–53 (2018) 1–15, http://dx.doi.org/10.1016/j.websem.2018.07.002.

[98] Z. Yang, A. Einolghozati, H. Inan, K. Diedrick, A. Fan, P. Donmez, S. Gupta, Improving text-to-text pre-trained models for the graph-to-text task, in: WebNLG Workshop, 2020, pp. 107–116.

[99] Y. Zhu, J. Wan, Z. Zhou, L. Chen, L. Qiu, W. Zhang, X. Jiang, Y. Yu, Triple-to-text: Converting RDF triples into high-quality natural languages via optimizing an inverse KL divergence, in: SIGIR, 2019, pp. 455–464, http://dx.doi.org/10.1145/3331184.3331232.

[100] G. Angeli, M.J. Johnson Premkumar, C.D. Manning, Leveraging linguistic structure for open domain information extraction, in: ACL-IJCNLP, 2015, pp. 344–354, http://dx.doi.org/10.3115/v1/P15-1034.

[101] K. Clark, C.D. Manning, Deep reinforcement learning for mention-ranking coreference models, in: EMNLP, 2016, pp. 2256–2262, http://dx.doi.org/10.18653/v1/D16-1245.

[102] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The stanford coreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60, http://dx.doi.org/10.3115/v1/P14-5010.

[103] Y. Matsuo, M. Ishizuka, Keyword extraction from a document using word co-occurrence statistical information, Trans. JSAI 17 (3) (2002) 217–223, http://dx.doi.org/10.1527/tjsai.17.217.

[104] J.D. Cohen, Highlights: Language- and domain-independent automatic indexing terms for abstracting, J. Am. Soc. Inf. Sci. 46 (3) (1995) 162–174, http://dx.doi.org/10.5555/205361.205362.

[105] H.P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM J. Res. Dev. 1 (4) (1957) 309–317, http://dx.doi.org/10.1147/rd.14.0309.

[106] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the First Instructional Conference on Machine Learning, Vol. 242, 2003, pp. 29–48.

[107] R. Campos, V. Mangaravite, A. Pasquali, A.M. Jorge, C. Nunes, A. Jatowt, YAKE! Keyword extraction from single documents using multiple local features, Inform. Sci. 509 (2020) 257–289, http://dx.doi.org/10.1016/j.ins.2019.09.013.

[108] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, M. Jaggi, Simple unsupervised keyphrase extraction using sentence embeddings, in: CoNLL, 2018, pp. 221–229, http://dx.doi.org/10.18653/v1/K18-1022.

[109] X. Liang, S. Wu, M. Li, Z. Li, Unsupervised keyphrase extraction by jointly modeling local and global context, in: EMNLP, 2021, pp. 155–164, http://dx.doi.org/10.18653/v1/2021.emnlp-main.14.

[110] J. Ye, R. Cai, T. Gui, Q. Zhang, Heterogeneous graph neural networks for keyphrase generation, in: EMNLP, 2021, pp. 2705–2715, http://dx.doi.org/10.18653/v1/2021.emnlp-main.213.

[111] M. Okamoto, M. Kikuchi, Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries, in: Asia Information Retrieval Symposium, 2009, pp. 181–192, http://dx.doi.org/10.1007/978-3-642-04769-5_16.

[112] M. Liu, Y. Liu, L. Xiang, X. Chen, Q. Yang, Extracting key entities and significant events from online daily news, in: IDEAL, 2008, pp. 201–209, http://dx.doi.org/10.1007/978-3-540-88906-9_26.

[113] W. Wang, Y. Ning, H. Rangwala, N. Ramakrishnan, A multiple instance learning framework for identifying key sentences and detecting events, in: CIKM, 2016, pp. 509–518, http://dx.doi.org/10.1145/2983323.2983821.

[114] D. Zhou, X. Zhang, Y. He, Event extraction from Twitter using non-parametric Bayesian mixture model with word embeddings, in: EACL, 2017, pp. 808–817, http://dx.doi.org/10.18653/v1/e17-1076.

[115] S.K. Bharti, K.S. Babu, Automatic keyword extraction for text summarization: A survey, 2017, http://dx.doi.org/10.48550/arXiv.1704.03242, arXiv preprint arXiv:1704.03242.

[116] F. Hill, A. Bordes, S. Chopra, J. Weston, The goldilocks principle: Reading children's books with explicit memory representations, in: ICLR, 2016, http://dx.doi.org/10.48550/arxiv.1511.02301.

[117] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, et al., Visual storytelling, in: NAACL, 2016, pp. 1233–1239, http://dx.doi.org/10.18653/v1/N16-1147.

[118] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, J. Allen, A corpus and cloze evaluation for deeper understanding of commonsense stories, in: NAACL, 2016, pp. 839–849, http://dx.doi.org/10.18653/v1/N16-1098.

[119] A. Fan, M. Lewis, Y. Dauphin, Hierarchical neural story generation, in: ACL, 2018, pp. 889–898, http://dx.doi.org/10.18653/v1/P18-1082.

[120] P. Ammanabrolu, W. Cheung, D. Tu, W. Broniec, M. Riedl, Bringing stories alive: Generating interactive fiction worlds, in: AIIDE, Vol. 16, 2020, pp. 3–9, http://dx.doi.org/10.48550/arXiv.2001.10161.

[121] A. Nader, W. Shufan, W. Josh, H. Stephen, P. Nanyun, I. Mohit, STORIUM: A dataset and evaluation platform for story generation, in: EMNLP, 2020, pp. 6470–6484, http://dx.doi.org/10.18653/v1/2020.emnlp-main.525.

[122] M. Sap, E. Horvitz, Y. Choi, N.A. Smith, J. Pennebaker, Recollection versus imagination: Exploring human memory and cognition via neural language models, in: ACL, 2020, pp. 1970–1978, http://dx.doi.org/10.18653/v1/2020.acl-main.178.

[123] M. Chen, K. Gimpel, TVRecap: A dataset for generating stories with character descriptions, 2021, http://dx.doi.org/10.48550/arxiv.2109.08833, arXiv.

[124] F. Brahman, M. Huang, O. Tafjord, C. Zhao, M. Sachan, S. Chaturvedi, "Let your characters tell their story": A dataset for character-centric narrative understanding, in: EMNLP Findings, 2021, pp. 1734–1752, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.150.

[125] J. Guan, Z. Liu, M. Huang, A corpus for understanding and generating moral stories, in: NAACL, 2022, pp. 5069–5087, http://dx.doi.org/10.18653/v1/2022.naacl-main.374.

[126] J.D. Hwang, C. Bhagavatula, R.L. Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs, in: AAAI, Vol. 35, 2021, pp. 6384–6392, http://dx.doi.org/10.48550/arXiv.2010.05953.

[127] N. Mostafazadeh, A. Kalyanpur, L. Moon, D. Buchanan, L. Berkowitz, O. Biran, J. Chu-Carroll, GLUCOSE: GeneraLized and COntextualized story explanations, in: EMNLP, 2020, pp. 4569–4586, http://dx.doi.org/10.18653/v1/2020.emnlp-main.370.

[128] H. Zhang, D. Khashabi, Y. Song, D. Roth, Transomcs: From linguistic graphs to commonsense knowledge, in: IJCAI, 2020, pp. 4004–4010, http://dx.doi.org/10.24963/ijcai.2020/554.

[129] B. Dalvi Mishra, N. Tandon, P. Clark, Domain-targeted, high precision knowledge extraction, TACL 5 (2017) 233–246, http://dx.doi.org/10.1162/tacl_a_00058.

[130] H. Zhang, X. Liu, H. Pan, Y. Song, C.W.-K. Leung, ASER: A large-scale eventuality knowledge graph, in: WWW, 2020, pp. 201–211, http://dx.doi.org/10.1145/3366423.3380107.

[131] P.R. Kingsbury, M. Palmer, From TreeBank to PropBank, in: LREC, 2002, pp. 1989–1993.

[132] C.F. Baker, C.J. Fillmore, J.B. Lowe, The berkeley framenet project, in: ACL-COLING, 1998, pp. 86–90, http://dx.doi.org/10.3115/980845.980860.

[133] L.R. Bahl, F. Jelinek, R.L. Mercer, A maximum likelihood approach to continuous speech recognition, TPAMI 5 (2) (1983) 179–190, http://dx.doi.org/10.1109/TPAMI.1983.4767370.

[134] Z. Shao, M. Huang, J. Wen, W. Xu, X. Zhu, Long and diverse text generation with planning-based hierarchical variational model, in: EMNLP, 2019, pp. 3255–3266, http://dx.doi.org/10.18653/v1/D19-1321.

[135] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, in: ACL, 2016, pp. 110–119, http://dx.doi.org/10.18653/v1/N16-1014.

[136] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in: ACL, 2002, pp. 311–318, http://dx.doi.org/10.3115/1073083.1073135.

[137] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81, URL https://aclanthology.org/W04-1013.

[138] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: ACL Workshop, 2005, pp. 65–72, URL https://aclanthology.org/W05-0909.

[139] R. Vedantam, C.L. Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: CVPR, 2015, pp. 4566–4575, http://dx.doi.org/10.1109/CVPR.2015.7299087.

[140] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: ICLR, 2020, http://dx.doi.org/10.48550/arXiv.1904.09675.

[141] J. Guan, M. Huang, UNION: An unreferenced metric for evaluating open-ended story generation, in: EMNLP, 2020, pp. 9157–9166, http://dx.doi.org/10.18653/v1/2020.emnlp-main.736.

[142] K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, Z. Harchaoui, Mauve: Measuring the gap between neural text and human text using divergence frontiers, Adv. Neural Inf. Process. Syst. 34 (2021) 4816–4828, http://dx.doi.org/10.48550/arXiv.2102.01454.

[143] M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, L. Charlin, Language GANs falling short, in: ICLR, 2020, http://dx.doi.org/10.48550/arXiv.1811.02549.

[144] S. Gretz, Y. Bilu, E. Cohen-Karlik, N. Slonim, The workweek is the best time to start a family – a study of GPT-2 based claim generation, in: EMNLP Findings, 2020, pp. 528–544, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.47.

[145] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthc. 3 (1) (2021) http://dx.doi.org/10.1145/3458754.

[146] Z. Zheng, X.-Z. Lu, K.-Y. Chen, Y.-C. Zhou, J.-R. Lin, Pretrained domain-specific language model for natural language processing tasks in the AEC domain, Comput. Ind. 142 (2022) 103733, http://dx.doi.org/10.1016/j.compind.2022.103733.

[147] Y. Yang, M.C.S. Uy, A. Huang, Finbert: A pretrained language model for financial communications, 2020, http://dx.doi.org/10.48550/arXiv.2006.08097, arXiv preprint arXiv:2006.08097.

[148] X. Peng, K. Xie, A. Alabdulkarim, H. Kayam, S. Dani, M.O. Riedl, Guiding neural story generation with reader models, 2021, http://dx.doi.org/10.48550/arXiv.2112.08596, arXiv.

[149] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, in: NeurIPS, 2019, http://dx.doi.org/10.5555/3454287.3455099.

[150] K. Jones, J.R. Nurse, S. Li, Are you robert or RoBERTa? Deceiving online authorship attribution models using neural text generators, in: ICWSM, Vol. 16, 2022, pp. 429–440, http://dx.doi.org/10.48550/arXiv.2203.09813.

[151] A. Uchendu, T. Le, K. Shu, D. Lee, Authorship attribution for neural text generation, in: EMNLP, 2020, pp. 8384–8395, http://dx.doi.org/10.18653/v1/2020.emnlp-main.673.

[152] D. Pascual, B. Egressy, C. Meister, R. Cotterell, R. Wattenhofer, A plug-and-play method for controlled text generation, in: EMNLP Findings, 2021, pp. 3346–3361, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.334.

[153] B. Li, L.J. Martin, C. Callison-Burch, CIS2: A simplified commonsense inference evaluation for story prose, in: ACL Workshop, 2022, http://dx.doi.org/10.48550/arXiv.2202.07880.

[154] X.L. Li, A. Kuncoro, J. Hoffmann, C.d.M. d'Autume, P. Blunsom, A. Nematzadeh, A systematic investigation of commonsense knowledge in large language models, in: EMNLP, 2022, http://dx.doi.org/10.48550/arXiv.2111.00607.

[155] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, J. Pineau, How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2122–2132, http://dx.doi.org/10.18653/v1/D16-1230.

[156] M. Sagarkar, J. Wieting, L. Tu, K. Gimpel, Quality signals in generated stories, in: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, 2018, pp. 192–202, http://dx.doi.org/10.18653/v1/S18-2024.

[157] A. McCabe, C. Peterson, What makes a good story, J. Psycholinguistic Res. 13 (6) (1984) 457–480, http://dx.doi.org/10.1007/BF01068179.

[158] R. Lowe, M. Noseworthy, I.V. Serban, N. Angelard-Gontier, Y. Bengio, J. Pineau, Towards an automatic turing test: Learning to evaluate dialogue responses, in: ACL, 2017, pp. 1116–1126, http://dx.doi.org/10.18653/v1/P17-1103.

[159] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7881–7892, http://dx.doi.org/10.18653/v1/2020.acl-main.704.

[160] F.R. Chang, Mental processes in reading: A methodological review, Read. Res. Q. (1983) 216–230, http://dx.doi.org/10.2307/747518.

[161] R.A. Mar, K. Oatley, M. Djikic, J. Mullin, Emotion and narrative fiction: Interactive influences before, during, and after reading, Cogn. Emot. 25 (5) (2011) 818–833, http://dx.doi.org/10.1080/02699931.2010.515151.

[162] J. Guan, Z. Zhang, Z. Feng, Z. Liu, W. Ding, X. Mao, C. Fan, M. Huang, OpenMEVA: A benchmark for evaluating open-ended story generation metrics, in: ACL-IJCNLP, 2021, pp. 6394–6407, http://dx.doi.org/10.18653/v1/2021.acl-long.500.

[163] Z. Xie, M. Li, T. Cohn, J.H. Lau, DeltaScore: Evaluating story generation with differentiating perturbations, 2023, http://dx.doi.org/10.48550/arXiv.2303.08991, arXiv.

[164] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: ICML, Vol. 139, 2021, pp. 8821–8831, http://dx.doi.org/10.48550/arXiv.2102.12092.

[165] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with CLIP latents, 2022, http://dx.doi.org/10.48550/arXiv.2204.06125, arXiv.

[166] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, 2022, pp. 10684–10695, http://dx.doi.org/10.48550/arXiv.2112.10752.

[167] D. Joshi, J.Z. Wang, J. Li, The story picturing engine—A system for automatic text illustration, ACM Trans. Multimedia Comput. Commun. Appl. 2 (1) (2006) 68–89, http://dx.doi.org/10.1145/1126004.1126008.

[168] S. Nag Chowdhury, W. Cheng, G. de Melo, S. Razniewski, G. Weikum, Illustrate your story: Enriching text with images, in: WSDM, 2020, pp. 849–852, http://dx.doi.org/10.1145/3336191.3371866.

[169] J. Zakraoui, M. Saleh, A. Ja'am, et al., Text-to-picture tools, systems, and approaches: a survey, Multimedia Tools Appl. 78 (16) (2019) 22833–22859, http://dx.doi.org/10.1007/s11042-019-7541-4.

[170] S. Chen, B. Liu, J. Fu, R. Song, Q. Jin, P. Lin, X. Qi, C. Wang, J. Zhou, Neural storyboard artist: Visualizing stories with coherent image sequences, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2236–2244, http://dx.doi.org/10.1145/3343031.3350571.

[171] Y. Chen, R. Li, B. Shi, P. Liu, M. Si, Visual story generation based on emotion and keywords, 2023, http://dx.doi.org/10.48550/arXiv.2301.02777, arXiv.

[172] X. Hong, A. Sayeed, K. Mehra, V. Demberg, B. Schiele, Visual writing prompts: Character-grounded story generation with curated image sequences, 2023, http://dx.doi.org/10.48550/arXiv.2301.08571, arXiv.

[173] W.F. Brewer, E.H. Lichtenstein, Stories are to entertain: A structural-affect theory of stories, J. Pragmatics 6 (5–6) (1982) 473–486, http://dx.doi.org/10.1016/0378-2166(82)90021-2.

[174] R. Han, H. Chen, Y. Tian, N. Peng, Go back in time: Generating flashbacks in stories with event temporal prompts, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2022, pp. 1450–1470, http://dx.doi.org/10.18653/v1/2022.naacl-main.104.

[175] S. Goldfarb-Tarrant, H. Feng, N. Peng, Plan, write, and revise: An interactive system for open-domain story generation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 89–97, http://dx.doi.org/10.18653/v1/N19-4016.

[176] F. Brahman, A. Petrusca, S. Chaturvedi, Cue me in: Content-inducing approaches to interactive story generation, in: ACL-IJCAI, 2020, pp. 588–597.

[177] P. Mirowski, K.W. Mathewson, J. Pittman, R. Evans, Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals, 2022, http://dx.doi.org/10.48550/arXiv.2209.14958, arXiv.

[178] A. Dirik, H. Donmez, P. Yanardag, Controlled cue generation for play scripts, 2021, arXiv preprint arXiv:2112.06953.

[179] P. Yanardag, M. Cebrian, I. Rahwan, Shelley: A crowd-sourced collaborative horror writer, in: Creativity and Cognition, 2021, pp. 1–8, http://dx.doi.org/10.1145/3450741.3465251.

[180] K. Yang, Y. Tian, N. Peng, D. Klein, Re3: Generating longer stories with recursive reprompting and revision, in: ICLR, 2022, http://dx.doi.org/10.48550/arXiv.2210.06774.

Yuxin Wang received her M.Eng and B.S degrees in Computer Science from Nanjing University, China, in 2023 and 2019, respectively. In 2022, she served as an intern at Microsoft Research Asia. Her research interests include knowledge graphs representation learning, entity alignment, and text generation.

Jieru Lin received the B.Eng degree in July 2021 from the Department of Computer Science, Harbin Institute of Technology, Harbin, China, where she has been working towards the Ph.D. degree, since 2021. Her research interests contain design intelligence, layout generation and natural language processing.

Zhiwei Yu received the Master of Natural Science degree in 2020, from Peking University, China. She is currently working as a researcher at Microsoft. Her research interests including information extraction, knowledge computing, and design intelligence.

Wei Hu is an associate professor at State Key Laboratory for Novel Software Technology, National Institute of Healthcare Data Science, Nanjing University, Nanjing, China. He received his PhD degree in Computer Software and Theory in 2009, and BS degree in Computer Science and Technology in 2005, both from Southeast University, China. His main research interests include knowledge graph, data integration and intelligent software.

Börje F. Karlsson received his PhD from PUC-Rio in 2010. He is current Member of Technical Staff at the Beijing Academy of Artificial Intelligence (BAAI), Beijing, China. He has also served on the program committee for major conferences, including AAAI, ACL, IJCAI, ACM MobileSoft, ICEC, etc. His interests include machine learning systems, multi-modality reasoning, intelligent agents, and digital storytelling.