Proceedings of the 2023 26th International Conference on Computer Supported Cooperative Work in Design

Logic Event Graph Enhanced Narrative Generation

Caiyan Li[1], Lizhen Cui[1], Yonghui Xu[2], Ning Liu[1]
[1]School of Software, Shandong University, Jinan, China
[2]Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR) & Software School, Jinan, China
[clz, liun21cs]@sdu.edu.cn, cyanli@mail.sdu.edu.cn, xu.yonghui@hotmail.com

2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD) | 979-8-3503-3168-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSCWD57460.2023.10152623

***Abstract*—Narrative generation aims at producing fluent long texts from input data, which is widely used in event prediction, story generation and other fields. Previous work mainly relied on text data, picture data, knowledge graphs, etc. Unfortunately, these static data only focused on entities and their relationships, cannot reflect the dynamic development of events in the real world. In addition, because the language model itself does not have the ability of logical reasoning, the generated results often lack logical reasoning. To address these challenges, several studies incorporate logic event graphs into the task of narrative generation. The Logic Event Graph(LEG) is essentially an event logic knowledge base, which describes the event evolution law and development modes. With the help of the LEG, the narrative generation system will acquire the human "reasoning" ability, and generate narratives that better reflect the objective evolution laws of human behaviors and events. While significant research has been conducted in these areas, gaps in the surveying literature still exist. In this survey,we—(1)analyze the context of narrative generation; (2)discuss the advantages of joining the LEG in narrative generate; (3)review common strategies in narrative generation; (4)summarize the evaluation metricse of narrative generate, and the advantages and disadvantages of each evaluation metrics are discussed; (5)present current challenges and future research directions. This survey can guide researchers and practitioners to understand the research progress in these areas.**

***Index Terms*—Narrative Generation; Logic Event Graph; Natural Language Processing.**

I. Introduction

Narrative generation is the process of training machines to master the human ability to create text, which has become one of the most critical and challenging tasks in natural language processing [18]. So far, this technology has been widely used in various fields, and many companies at home and abroad have begun to use AI article generators to create contents. For example, companies such as ContentGenius, China's Jiaoyou, iWriter, and Automated Insights in the United States are pioneers in using this technology. Among them, hundreds of millions of reports have been written by Automated Insights's 'language experts' . According to the different forms of input data, narrative generation can be divided into data-to-text generation and text-to-text generation. The data can be of various types, such as semantic data, image data, and knowledge graphs [32].

However, traditional text and data only focus on the connections between entities while ignoring the existence of many events in the world and the dynamic timing knowledge transmitted between events, which reflects the sequence of events in time and space—evolution laws and patterns. Therefore, the Logic Event Graph comes into being to obtain a more "advanced" knowledge representation.

In 2019, [10] proposed the concept of an Logic Event Graph (LEG) for the first time. In graph structure, LEG is directed and cyclic, in which nodes represent events, and directed edges represent the logical relationships between events, such as sequence, cause and effect, conditions, and timing [11]. Compared with entities, taking events as the basic unit of knowledge can better reflect the inside of the objective world, especially learning dynamics [37]. From the cognitive psychology perspective, events align with human understanding and thinking habits. With the help of LEG, machines can generate a particular "logical reasoning" ability, which is very difficult and meaningful in narrative generation tasks. For example, humans can easily understand the common sense knowledge of "buying a fishing rod" for "fishing". Still, it is complicated for machines to understand and master a large amount of such knowledge. Specifically, because the language models do not have the ability of logical reasoning, language models must train a lot, and the training data must have event logic (such as LEG). After training, the machine will perform well in some generation tasks involving reasonings and predictions [20]. For example, after the machine knows the event "marriage", it can automatically generate descriptions of events such as "purchasing furniture" and "buying a new house". It can also cause a brief introduction of the decoration company based on the "buying a house" event and recommend it to users.

Even though logic event graph-enhanced narrative generation has a wide range of application scenarios in various fields [26], narrative generation systems still need to be able to objectively reproduce naturally occurring events, mainly due to the following two aspects. The first is that generative systems lack logical reasoning, and most existing narrative generation is trained with maximum likelihood estimation from large-scale dialogue data [1]. However, events in the real world are complex and changeable. It is difficult for existing generative models to understand the cause and effect of the context deeply, but only the memory of specific patterns in the training corpus. Only when the machine understands the common sense that "buy a fishing rod" is for "fishing" and "buy a ticket" before "watching a movie" can the narrative generation system generate more appropriate text according to different generation tasks. The second is that the evaluation indicators need to be completed, making it challenging to improve the model's performance. Narrative generation is different from general text generation tasks. It needs to pay attention to events' integrity, relevance, and logic from a macro perspective. Still, traditional evaluation indicators often only focus on character-level evaluation, which is also a hindrance to narrative generation—a big reason.

Although some surveys have reviewed the development of narrative generation, including the fusion of knowledge [41] and the improvement of language models [29], the research on narrative generation enhanced by LEG still needs to be completed. Integrating LEG into narrative generation tasks can improve the model's commonsense reasoning and prediction capabilities and generate narratives that are more in line with the development of actual events. Therefore, it is necessary to delve into narrative generation augmented by LEG. Consequently, we conduct a comprehensive investigation of the study.

The arrangement of this article is as follows: First, the definition and construction of LEG are introduced in 2 chapter, and then the narrative generation strategy based on LEG enhancement is presented from different perspectives in 3 chapter, the narrative generation is summarized in 4 chapter Systematic Evaluation Indicators, including the advantages and disadvantages of these evaluation indicators; finally, in Chapter 5, current challenges and future research directions are introduced.

II. Preliminaries

A. Defining Logic Event Graph Narrative

LEG narrative generation task aims to generate highly relevant text for the input graph. This task requires simultaneously encoding the structural and entity information of the graph and efficiently exploiting this information during the decoding process [30].

LEG narrative generation typically requires a series of triples denoted as $G = \{< h, r, t > | h, t \in E, r \in R\}$, with E and R representing entities and relations between entities, respectively. A triple $< h, r, t >$ represents the corresponding entity that connects the head node h and the tail node t with the relation r. [18] described graph narrative generation as, first given a vocabulary V, the purpose of graph narrative generation is to generate natural language text $Y = \{y_1, y_2, \ldots, y_k\}$, Graph G's structure and content information are described by Y . For example, in Figure 1, a simple triple "['2004 AFC Asian Cup Final','winner','Japan national football team']", can be expressed linguistically as "The winner of the 2004 AFC Asian Cup Final is Japan national football team".

B. Logic Event Graph Construction

Constructing LEG is the basis for realizing our tasks. The construction of the logic event graph includes four stages: event extraction, event relationship extraction, event classification, and graph completion. Event extraction and event relationship extraction are the two core steps. i)Event extraction: Event extraction aims to extract events of interest to users from massive data and express them in structured information. Event extraction is roughly divided into two steps: event discovery and event element extraction. The former is mainly to find event triggers from the text (such as the time, place, role, process, and result of the event), the latter is to identify the event elements from the text and judge the role played by the factors. ii)event relationship extraction: The event relationship mainly includes the coreference, temporal, and causality relationships. Coreference refers to events that represent the same target. For example, "Disney's CEO denies that Apple will acquire the company" and "Disney and Apple have not signed an acquisition agreement" represent the same event. The coreference generally needs to be resolved. Feature learning-based methods and deep learning-based methods are two directions of relation extraction. Early event relation extraction methods relied heavily on artificially designed syntactic and semantic features. Machine learning models such as naive Bayes, maximum entropy, or support Vector machines(SVM) are used as classifiers to identify relationships [7], [4], [31]. But this method is too complex and highly dependent on specific domains. Recent studies have found that methods based on deep learning can more accurately extract event relationships. Therefore, mainstream event relationship extraction mainly relies on deep learning methods like LSTM, CNN, and RNN. iii)event classification and graph completion: In the previous steps, we obtained a large number of scattered events, and by connecting semantically related events, a preliminary event graph can be obtained. However, due to the incompleteness of the source text and the loss of information during the extraction process, it is necessary to use graph completion technology to fill in the event graph. Completion techniques mainly include prediction and classification methods. The former predicts based on known elements, while the latter involves learning candidate elements and selecting the element with the highest score.

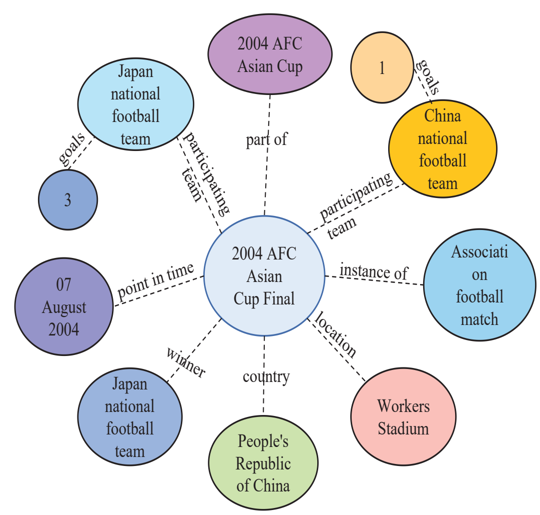

Fig. 1. An example logic event graph to narrative generation. Event narration: "The **2004 AFC Asian Cup Final** was an **association football match** that took place on **07 August 2004** at the **Workers Stadium in Beijing**, **People's Republic of China**, to determine the winner of the **2004 AFC Asian Cup**. **Japan national football team** defeated **China national football team 3 : 1**."

LEG generation is a data-to-text generative task that increases the accessibility of structured knowledge. Previous work has shown that pre-trained language models perform exceptionally well on a sizeable task-specific training [22]. However, since no suitable method has been found to integrate these structured data into model learning tasks perfectly, some problems have arisen. For example, loss of graph structure information and text degradation. The loss of graph structure information means that the language model may mess up or even lose the core structure information of the input graph when generating output; Text degradation mainly refers to problems such as repetitive results and lack of sentence diversity. The emergence of these problems is directly related to the design of the generation method. Next, we will introduce the generation strategy in Chapter 3.

III. Logic Event Graph Generation Strategies

In this subsection, we summarize the strategies of event narrative generation in four aspects: Common sense event representation with knowledge integration, Structure-assisted logical representation of events, Complete event generation with knowledge completion, and Diverse expressions of multi-party regulated events, respectively.

1) Complete graph enhanced: The completeness of the event narrative refers to the time and place of an event, the characters involved, and the event's cause, process, and outcome. Due to the realities of vast amounts of data and sloppy organization, traditional narrative generation based on language models struggles to depict a complete history of events. As a significant event information carrier, LEG has extracted and reasonably expressed event-related information, making it appropriate for constructing a complete event narrative. Unfortunately, because the current LEG is far from comprehensive, it needs descriptions of related events and their links. Knowledge Graph Completion approaches have helped to close this gap. The general approach to knowledge completion uses observed facts to model and infer connectivity patterns in the knowledge graph [36]. Learning low-dimensional representations of things and relationships suggested by missing connections has been the subject of extensive research (also known as knowledge graph embeddings).

2) Common sense enhanced: Because neural network architecture models need to gain common sense capabilities, they fall short in event narration tasks that require logical thinking and cognition. While the knowledge graph's rich common sense concepts can effectively address this issue. Graph networks are widely used for knowledge integration in graph structures, where dependencies between nodes provide access to full-text information. The common knowledge graphs are knowledge graph(KG) [18] Abstract Meaning Representation(AMR) graph [30], graph-to-sequence(Graph2Seq) [3] and so on; Concerning how to train the model to represent knowledge correctly, The attention mechanism is commonly employed in the creation of information fusion narratives. Its main idea is that, given a Graph sequence description, the generated text is of higher quality by giving higher weight to the keywords in a graph [40]; In recent years, as the power of computing platforms keeps increasing,pre-trained language models trained on large-scale datasets have been well received in text generation. Much research has been conducted on incorporating knowledge into pre-trained language models (PLM). Table 1 shows some language generation models, among which Rewriter does not incorporate knowledge representation, but Graphwriter incorporate; they did not pre-train on a large dataset; Next is the pre-trained language model; Graformer is generated with global information incorporated. A semantic aggregation module is added to Joint, enhancing the model's ability to learn graph structure. BART [22] and T5 [27] are two typical pre-trained language models; It can be seen that these language generation models have a substantial improvement in generation after incorporating knowledge representation, pre-training, and adding graph structure information.

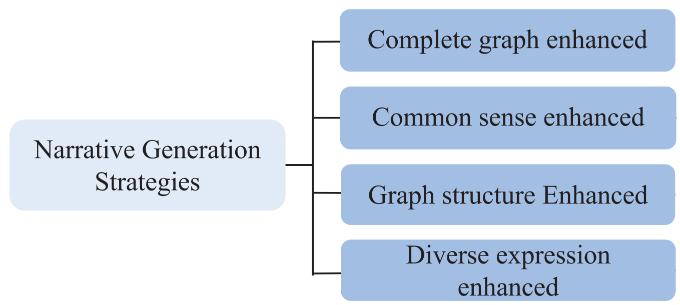

Fig. 2. Narrative Generation Strategies.

3) Graph structure Enhanced: When people write or express themselves, they organize the order of description based on the intrinsic connections of things or the process of people knowing things. However, the language model chooses the output based on probabilities; certain crucial concepts may be out of order, a common occurrence in graph narrative-generating tasks [1]. However, the order of these directed edges is significant because they represent the logical relationship between events. One option to overcome this challenge is to let the model learn the graph structure information, and there are three popular approaches: a.) Adding a semantic aggregation module to the encoder to preserve graph structure information, like JointGT [17]. b.) Some researchers are learning the target sentence and the corresponding graph structure in a decoder, such as CycleGT [30]; c.) Improving the model by reconstructing the graph structure on the output result and then calculating the loss with the original graph Learning ability of graph structure information [34]. As we can see in Table 1, adding this module to BART and T5 significantly improves the generation results.

4) Diverse expression enhanced: Narrative authoring aims to enable language models to produce grammatically correct, content-relevant, and semantically rich event narratives. Recent research has discovered that language models produce variable degrees of tedious repetition at the word, phrase, and sentence levels. To alleviate this problem, Goodfellow et al. introduced Generative Adversarial Network (GAN) into the text generation task to evaluate the model generation effect. The adversarial network's basic notion is to penalize the model's repeated generation behavior by constantly altering the game between the generator and the discriminator to guide different text generation. After that, a series of GAN variants for facilitating text diversity generation: seqGAN, RankGAN, and LeakGAN [39], [16], [12] were proposed successively. However, when the generated narratives' quality and diversity are examined, the GAN-based results are far inferior to the language model-based results. Following this, some research has indicated that an ineffective decoding approach causes duplication and that the decoding strategy can be changed by sampling from improbable vocabularies to reduce duplication [14]. Recent research has discovered that the distribution of token vectors learned by the model is anisotropic in semantic space, making it difficult to discern between various tokens. To address this issue, they incorporate contrast learning into the model training phase and contrast search into the decoding phase to improve vector isotropy across the representation space [35].

TABLE I
REWRITER AND GRAPHWRITER PERFORMANCE ON AGENDA; KGPT,T5,BART AND JOINT(BART) PERFORMANCE ON WEBNLG.

	Blu	Met	Rou	Knl	pre-tr	strc
Rewriter [23]	1.05	8.38	-	×	×	×
GraphWriter [18]	15.3	18.8	22.03	✓	×	×
Graformer [28]	61.15	43.38	-	✓	✓	×
KGPT [8]	61.44	46.30	74.57	✓	✓	×
T5 [27]	64.42	46.58	74.77	✓	✓	×
BART [22]	64.55	46.51	75.13	✓	✓	×
Joint(BART) [17]	65.92	47.15	74.77	✓	✓	✓

"Blu"means BLUE,"Met" means METEOR,"Rou" means Rouge,"Knl"means Knowledge,"pre-tr"means pre-trained,"strc"means Graph structure

Evaluation metrics are the basis for enhancing a system. The model would have no idea where to improve if it did not include evaluation measures. As a result, evaluating the quality of text created by the model, such as whether the generated results are loyal to the input and whether the utterance is fluent, grammatical, and natural-looking, is a crucial component of the text generation task [19].

IV. EVALUATION

Existing text evaluation metrics can be divided into two categories: manual evaluation and automatic evaluation. The results of the manual evaluation are highly accurate but inefficient and unsuitable for large-scale data sets. Automatic evaluation has advantages in terms of efficiency, but the evaluation content is relatively single and conflicting [13], so focusing on individual automatic evaluation metrics may not evaluate the model's performance. We summarize the metrics common to text generation in recent years and propose innovative new evaluation metrics for logic event graph-based narrative generation, which we hope will help researchers in this field.

A. Objective evaluation

Most general evaluations come from text-to-text generation tasks, such as dialogue generation and translation domains. Most compare the word overlap rate between the generated text and the reference text to gauge the model's performance. The classical methods are BLEU, METEOR, and ROUGE, among which BLEU and METEOR are commonly used for machine translation tasks and ROUGE is commonly used for automatic text summarization.

1) BLEU: The BLEU metric was proposed in 2002 [24], and this metric was one of the first metrics used to compare the similarity between candidate and reference sentences. The basic idea is to assess the degree of overlap between the n-grams in the candidate and reference translations, with the more significant overlap indicating higher translation quality.

2) METEOR: Unlike BLEU, METEOR supports matching not only between identical words in two strings but also simple morphological variants of each other (i.e., they have the same stem) and synonyms of each other [2] . In addition, METEOR considers the accuracy and recall based on the entire corpus to obtain a combined result.

3) ROUGE: ROUGE's full name is Recall-Oriented Understudy for Gisting Evaluation, which looks at how many n-grams from the reference translation appear in the output. However, because ROUGE only focuses on solving the problem of low recall, it is not more informative in measuring the fluency of the generated results in terms of accuracy [21].

4) chrF++: The significant difference between CHRF and BLEU-like metrics is that BLEU is word-level, and CHRF++ is concerned with the character-level generation based on character n-gram calculation while considering morpheme overlapping [25].

The above results show that most automatic evaluation metrics are still for text-based input tasks, and either metric has certain drawbacks. Methods based on word overlap rates such as BLEU, ROUGE, and other metrics are only sensitive to lexical changes. They do not adequately reward semantic or syntactic changes for a given reference. As a result, they have been demonstrated to have poor correlations with human evaluation, particularly when all of the systems being examined have equal degrees of accuracy [33]. Evaluation metrics based on word overlap do not consider the particularity of graph structures. To remedy this shortfall, [38]proposed some evaluation Metrics specifically for evaluating data-to-text generation. RG, CS, and CO are proposed.

5) RG: RG refers to relation generation, which measures the ability of the system to generate text containing factual (i.e., correct) records by comparing the extracted triples with the content in Gold text.

6) CS: CS refers to Content Selection, which measures how well the generated text matches the gold text in some of the selected generated content.

7) CO: CO means Content Ordering, measures the distance between extracted content and is used to measure how well the system sorts the content.

B. subjective evaluation

The subjective evaluation method is still the most dependable because of its correctness and the difficulty of text generation. And in developing new evaluation methods, subjective evaluation is considered the gold standard [6]. We summarize subjective evaluation in terms of Intrinsic Evaluation, Extrinsic Evaluation, and evaluators, depending on the subject of the evaluation.

1) Intrinsic Evaluation: The internal evaluation of the generated content mainly compares the content of the generated text with the content of the source text. It compares fluency, adequacy (whether the meaning of the original text is fully expressed), authenticity (the generated content is based on the source rather than the result of model fantasy), logic (whether the current action is consistent with the context), and then score each option to get a composite score.

2) Extrinsic Evaluation: Extrinsic evaluation is the most effective language model evaluation in terms of outcome, as it assesses the model's utility based on the generated text's performance on the downstream subtask. The results are judged to be improved by simple metrics, such as language recognition tasks.

3) Evaluators: By recruiting users, the technique participates in the evaluation of results. The model is changed in real-time to generate satisfactory results, and interaction between users and professional assessors is rewarded.

Subjective assessment is simple and easy to operate, and the results are relatively reliable. However, the results of the manual evaluation are highly imaged by subjective factors of evaluators, such as cultural background, professional knowledge, and cognitive ability. These individual differences may lead to inconsistent assessment results [15]; quality assurance is complex. Meanwhile, the manual evaluation's experimental results are frequently challenging to duplicate.

Most metrics are still in the blank stage, while evaluation metrics for the homogeneity and completeness of events in graph story production where the input is a graph structure are still in the blank stage. Precisely, homogeneity measures whether the narratives generated for the exact event map recount the same event; completeness refers to how many events are generated in the narrative for the same story. Equivalent to the accuracy and completeness rates in natural language. However, until now, these assessments could only be performed manually. In the future, it will be critical to developing a set of assessment metrics that includes a wide variety of evaluations of logic event graph structure and content. For event-based graph narrative generation tasks, it is crucial to judge the logic, coherence, and narrative integrity between the generated content contexts. However, current automatic evaluation metrics need to address this issue better and rely on efficient manual evaluation.

V. Challenges and Future Directions

Through our research, we found that the field of graph narrative generation is still in the exploratory stage, and there is still much room for development in the future. In this regard, we look forward to some future research directions.

1) Multi-source fusion: First of all, in terms of the extracted objects, there is a large amount of knowledge stored in various media such as pictures, text, speech, and videos in the real world. Currently, event knowledge is mainly obtained from text, and it is difficult to obtain multimodal knowledge that integrates speech, images, text, and videos. However, knowledge from different modalities can often complement or even promote each other. Secondly, although existing automatic knowledge acquisition ensures efficiency in terms of time, the accuracy of the extracted content is compromised. Therefore, how to combine multimodal information and improve the extraction method to obtain high-precision knowledge is an urgent problem to be solved in the development of future logic event graphs.

2) Robust Evaluation Metrics : As previously stated, text evaluation using logic event graph generation is difficult. Because the evaluation requirements of event narrative generation tasks differ from those of traditional generation tasks (such as translation, dialogue, and QA), the evaluation focus of the latter tends to focus on the overlap rate of words, which is relatively easy to achieve. However, event narratives generated based on an logic event graph need to achieve sentence fluency, coherent and reasonable paragraphs, exclusive events, etc., to ensure word accuracy. Although some researchers have proposed related theories, such as relation generation (RG) and content selection (CS) [38], the overall effect is not apparent. In the future, developing a set of metrics is imperative for the comprehensive and accurate evaluation of generated event narratives.

3) Eliminate prejudice: From the previous article, we can see that many researchers have done much work on ensuring the generation of high-quality graph narratives, and it has been verified that it is feasible. Nevertheless, little attention has been paid to the fact that the results can be biased. The use of pre-trained language models to generate biased results has recently been confirmed by some researchers [9], [5]. The reason can be explained that when the model is trained on a large data set, social bias (such as gender and race) will be encoded into the model. We know that these models depend on the training set's characteristics. In the generated results, There may also be biased results. This problem also exists in graph narratives. How to avoid biased graph narrative generation, is an issue that future research needs to pay attention to.

VI. Acknowledgement

This research is partially supported by the National Key R&D Program of China 2021YFF0900800, NSFC No.62202279, the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (No.2021CXGC010108), the Shandong Provincial Natural Science Foundation (No.ZR2022QF018), Shandong Provincial Outstanding Youth Science Foundation, the Fundamental Research Funds of Shandong University.

REFERENCES

[1] Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. Constrained decoding for neural nlg from compositional representations in task-oriented dialogue. *arXiv preprint arXiv:1906.07220*, 2019.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Daniel Beck, Gholamreza Haffari, and Trevor Cohn. Graph-to-sequence learning using gated graph neural networks. *arXiv preprint arXiv:1806.09835*, 2018.

[4] Eduardo Blanco, Nuria Castell, and Dan Moldovan. Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[5] Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *arXiv preprint arXiv:2203.13928*, 2022.

[6] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

[7] Nathanael Chambers, Shan Wang, and Dan Jurafsky. Classifying temporal relations between events. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 173–176, 2007.

[8] Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online, November 2020. Association for Computational Linguistics.

[9] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872, 2021.

[10] Xiao Ding, Zhongyang Li, Ting Liu, and Kuo Liao. Elg: an event logic graph. *arXiv preprint arXiv:1907.08015*, 2019.

[11] Saiping Guan, Xueqi Cheng, Long Bai, Fujun Zhang, Zixuan Li, Yutao Zeng, Xiaolong Jin, and Jiafeng Guo. What is event knowledge graph: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[12] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[13] Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*, 2019.

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

[15] David M Howcroft, Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, 2020.

[16] Felix Juefei-Xu, Rahul Dey, Vishnu Naresh Boddeti, and Marios Savvides. Rankgan: a maximum margin ranking gan for generating faces. In *Asian Conference on Computer Vision*, pages 3–18. Springer, 2018.

[17] Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. *arXiv preprint arXiv:2106.10502*, 2021.

[18] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*, 2019.

[19] Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*, 2022.

[20] Zhongyang Li, Xiao Ding, and Ting Liu. Constructing narrative event evolutionary graph for script event prediction. *arXiv preprint arXiv:1805.05081*, 2018.

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[22] Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6418–6425, 2021.

[23] Diego Marcheggiani and Laura Perez-Beltrachini. Deep graph convolutional encoders for structured data to text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 1–9, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[25] Maja Popović. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618, 2017.

[26] Zifeng Qi, Haiguang Chen, Jifeng Huang, Chenglei Yu, and Mingxing Liu. Logical events graphs based emotional text generation. In *Proceedings of the 10th International Conference on Computer and Communications Management*, pages 54–60, 2022.

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[28] Leonardo F. R. Ribeiro, Yue Zhang, Claire Gardent, and Iryna Gurevych. Modeling Global and Local Node Contexts for Text Generation from Knowledge Graphs. *Transactions of the Association for Computational Linguistics*, 8:589–604, 09 2020.

[29] Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*, 2020.

[30] Leonardo FR Ribeiro, Yue Zhang, and Iryna Gurevych. Structural adapters in pretrained language models for amr-to-text generation. *arXiv preprint arXiv:2103.09120*, 2021.

[31] Bryan Rink, Cosmin Adrian Bejan, and Sanda Harabagiu. Learning textual graph patterns to detect causal event relations. In *Twenty-Third International FLAIRS Conference*, 2010.

[32] Lee Schruben. Simulation modeling with event graphs. *Communications of the ACM*, 26(11):957–963, 1983.

[33] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

[34] Linfeng Song, Ante Wang, Jinsong Su, Yue Zhang, Kun Xu, Yubin Ge, and Dong Yu. Structural information preserving for graph-to-text generation. *arXiv preprint arXiv:2102.06749*, 2021.

[35] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. *arXiv preprint arXiv:2202.06417*, 2022.

[36] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

[37] Xiang Wei. A review of event knowledge graph construction techniques and applications. *Computers and Modernization*, 1(10), 2020.

[38] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.

[39] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[40] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*, 2020.

[41] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Computing Surveys (CSUR)*, 2022.