Generating Stories by Prompting Pre-trained Language Models

Paresh Pradhan

Research Proposal for
Master of Science in Machine Learning & Artificial Intelligence

Liverpool John Moores University & upGrad

November 2022

Abstract

Text generation problems have seen remarkable advancements with the advent of pre-trained language models (PLMs). These models can only influence a few broad characteristics of the output text, though. PLMs are unable to produce long-form stories as they are ignorant of the narrative structure. Recent studies on tale generation have made use of explicit content planning, which can result in stories with more logical event sequences. However, a shortage of training data makes it challenging to fine-tune PLMs. It is difficult to achieve precise control, even with fine-tuning. Therefore, building a model that can produce long-form stories is not simple. Recent prompt-based learning provides a potential answer for this problem. This thesis work proposes a method to use prompt-based learning to generate stories while maintaining fine-grained control.

Table of Contents

1. Background

In recent years, with the emergence of large pre-trained language models, the quality of machine-generated text has improved significantly (Rashkin et al., 2018; Radford et al., 2019; Zhang et al., 2019; Brown et al., 2020; Guan et al., 2020; Bakhtin et al., 2021). Today, models can generate text that is indistinguishable from human-written text (Clark et al., 2021).

Controlling the generation is still a challenge however, despite the fact that large-scale PLMs have demonstrated excellent capabilities in producing coherent and intelligible text (Keskar et al., 2019; Radford et al., 2019; Zellers et al., 2019). A more thorough examination of generated text reveals problems like topic drift and self-contradiction (Fan et al., 2019; Bisk et al., 2020; Gao et al., 2020a; Tan et al., 2020; Dou et al., 2021; Dziri et al., 2021). These flaws stand out, in particular, for open-ended text generation tasks that require a high level of coherence, such as story generation.

Stories generated using language models have shown to lack discourse coherence (Bosselut et al., 2018; Ji and Huang, 2021), global planning (Hua and Wang, 2020; Tan et al., 2020) and common-sense knowledge (Ji et al., 2020; Xu et al., 2020). While the individual sentences in a generated text seem logical and fluent, when put together, the overall story often does not make much sense (See et al., 2019; Goldfarb-Tarrant et al., 2020). In long-form text generation, sentences tend to repeat which leads to reduction in story quality (Yao et al., 2019).

To provide structure to the generation process, recent works have tried to used explicit content planning. The content plan comes in different forms. (Fan et al., 2018a) used prompts. (Xu et al., 2018; Yao et al., 2019) used keywords and key-phrases. (Fan et al., 2019) used semantic frames. (Sun et al., 2020) used summaries. To make use of these content plans, PLMs generally require fine-tuning on content-plan related data. A challenge with fine-tuning PLMs is that, in addition to needing training data, the model has a tendency to learn frequently occurring events from the content plan and derives common sense information from them (Fan et al., 2019). This leads to lack of variety in generated stories.

Another novel approach to address linguistic issues without fine-tuning is provided by the recently proposed prompt-based learning (Liu et al., 2021). In this framework, task-specific prompts can be used to address text-based problems. Researchers have shown that using prompts, PLMs can solve existing or new generation tasks without need for fine-tuning (Brown et al., 2020; Li and Liang, 2021).

Although prompt-based learning looks promising, there are still some challenges. Prompts are highly task-specific and are hard to transfer or reuse for new tasks (Gao et al., 2020b). Even for the same task the prompts may not work well for all instances in a large population (le Scao and Rush, 2021).

2. Related Work

Controllable story generation has been studied from different angles. Researchers have focused on controlling story generation using broad thematic elements such as sentiment, genre, style, topic, etc. (Hu et al., 2017; Shen et al., 2017; Zhao et al., 2018; Dathathri et al., 2019; Fang et al., 2019; Keskar et al., 2019). Some works have tried more fine-grained control using story-lines, story-plans and plots (Peng et al., 2018; Yao et al., 2019). These works were benchmarked using relatively short-text datasets such as the five-lines story dataset, ROCStories (Mostafazadeh et al., 2016). Later on, some works have tried to controllable story generation with long-form text (Fan et al., 2018a, 2019; Rashkin et al., 2020; Fang et al., 2021).

Similar to short story generation, researchers have tried using fine-grained control to drive long-form story generation as well. (Fang et al., 2021) proposed generation of story given an outline of story events/phrases. (Rashkin et al., 2020) presented a comparable technique with a specialized architecture and memory mechanism. (Sun et al., 2020) created an outline of the story by generating summaries for each segment of the story. Then each summary is extrapolated to generate the full story.

Most of the research in the field is based on fine-tuning transformer-based Pre-trained Language Models (PLM) (Vaswani et al., 2017; Devlin et al., 2018; Radford et al., 2019) with curated or generated datasets (Conneau and Lample, 2019; Dong et al., 2019; Keskar et al., 2019; Song et al., 2019). Because of its specialized architecture for unconditional text generation, GPT2 (Radford et al., 2019), in particular, has attracted much attention in this field (Mao et al., 2019; See et al., 2019; Ziegler et al., 2019). And lately, after the availability of its API, GPT3 (Brown et al., 2020) has seen increasing usage for text generation (Dou et al., 2021; Shakeri et al., 2021).

In the absence of sufficient data, fine-tuning PLMs is challenging (Chen et al., 2019; Li et al., 2021). To resolve that, researchers have tried Plug-and-Play methods to control story generation

without fine-tuning (Dathathri et al., 2019; Pascual et al., 2020, 2021; Lin and Riedl, 2021; Jin et al., 2022; Mori et al., 2022).

Prompt-based learning is another approach that does not require fine-tuning. Some works have used hand-crafted prompts for various generation tasks (Brown et al., 2020; Raffel et al., 2020; Zou et al., 2021). Others have tried to automatically generate discrete prompts (Gao et al., 2020b; Shin et al., 2020) and continuous prompts (Li and Liang, 2021; Liu et al., 2021). Some have tried to generate prompts for target task using source task (Su et al., 2021; Vu et al., 2021).

3. Research Questions

This thesis tries to answer the following questions:

1. The approaches for story generation with fine-grained control require fine-tuning of PLMs. Can these approaches be used with Prompt-based learning to generate stories in a Few-Shot manner without fine-tuning?

2. The previous methods largely use GPT2 as base model. Can using the latest generation GPT3 (or alternatives) improve the text generation capabilities?

3. Prompt-based learning has been used to generate text in few-shot manner. Can this be extended to story generation task?

4. Aim and Objectives

This work tries to explore the Few-shot capabilities of GPT3 for long-form controllable story generation task.

Objectives:

- To conduct a comprehensive review of available literature with regards to Long-form story generation, Prompt-learning and Few-Shot text generation.

- To explore the viability and then develop a method to generate short and long form stories using few-shot generation and prompting.

- To evaluate the generated stories using automated story generation evaluation metrics and compare the developed method against existing methods.

5. Significance of the Study

Story Generation is a field under active research. While short-form story generation has been studied extensively, long-form story generation is relatively under-explored. Although fine-tuning based approaches have been used in previous works, there is a lack of research in generating stories without fine-tuning.

This work tries to fill these gaps by adding to the existing literature, providing benchmarks and contributing code. This work also explores recent developments in Prompt-based learning and Few-Shot generation.

In terms of application, this work helps story writers write better stories in conjunction with AI. This can help writers get new ideas or get over the writer's block.

6. Scope of the Study

The scope of this thesis work is defined as follows:

- The thesis work is to be completed within 17 weeks after submission of research proposal.
- The experimentation will be conducted using open-source software and models.
- The experimentation will be conducted using publicly available GPU such as Google-Colab.
- Human evaluation of the generated story is not a part of this thesis work. The evaluation will only focus on automated metrics.

7. Research Methodology

This work focuses on the text generation and few-shot learning capabilities of PLMs. Given an outline as a control-mechanism, the model should generate a story conditioned on the outline.

7.1 Dataset Description

This work makes use of two standard story generation datasets:

- **ROCStories**: Introduced by (Mostafazadeh et al., 2016), this dataset contains ~98K 5-sentence long stories along with story titles. This dataset is widely used for short-form story generation tasks.

- **WritingPrompts**: Introduced by (Fan et al., 2018b), this dataset contains ~300K human-written stories along with the starting prompt used to write the story. These stories were collected from the Reddit, an online social media forum. These stories are long-form multi-paragraph stories, and hence useful for more complex task of long-form story generation.

7.2 Data Preparation

The proposed method requires sample pairs of outline-instance to paragraph. While paragraph text can be derived from the ROCStories and WritingPrompts datasets, there is no dataset of outlines readily available. Hence, the outlines need to extracted from the story datasets and then mapped to corresponding paragraph text. These outline-paragraph pairs can then be sampled during the few-shot inference.

The outline instances can take one of two forms:

- **Summary** – Here the outline instance is a short extractive summary of the paragraph. The paragraph is expanded from the summary. For the summary extraction, TextRank (Mihalcea and Tarau, 2004) is proposed to be used to extract the most informative sentence from the paragraph.

- **Keywords/Keyphrases** - Here the outline instance is a set of keywords and phrases that are present in the paragraph. The paragraph text is generated conditioned on these keywords/keyphrases. For the outline extraction, RAKE (Rose et al., 2010) is proposed to be used to extract keyphrases from the paragraph.

7.3 Algorithms & Techniques Description

7.3.1 Pre-trained Language Models (PLMs)

Pre-trained models originate from the idea of transfer learning. Transfer learning refers to the process of applying previously acquired knowledge to new tasks. Traditional transfer learning used large volume of annotated data points for supervised training. Pre-training with self-supervised learning on vast amounts of unlabelled data has emerged as the most popular transfer learning strategy in deep learning. Pre-training methods differ from other approaches in that they use unlabelled data for self-supervised training and can be used for a number of downstream tasks using fine-tuning or few-shot learning.

Language modelling, in NLP, refers to the task of predicting the next character/word/sentence in a text. Language models are trained in a self-supervised manner using large corpora of unstructured text. These models can then be used for a number of natural language tasks, such as question answering, text generation, and text classification.

Pre-trained language models combine the tasks of language modelling and transfer learning leading to the creation of large language models which can be fine-tuned for many downstream tasks. Some of the most well-known language models are:

- BERT (Devlin et al., 2018)
- GPT3 (Brown et al., 2020)

7.3.2 Few-Shot Learning (FSL)

With the aid of a small number of samples and previously acquired knowledge, humans can quickly recognise new classes in data. This is called meta-learning. Few-Shot Learning is a type

of meta-learning. In this method, to efficiently generalise to new (but related) tasks with a small number of instances during the meta-testing phase, a learner is taught on a number of related tasks during the meta-training phase. One effective approach towards solving Few-Shot Learning problems is to learn a common representation for numerous tasks and then train task-specific classifiers on top of this representation. FSL is a solution to the problem of traditional supervised learning methods requiring large quantities of labelled data for training.

7.3.3 Prompt-Learning

Prompt-based learning is a new class of techniques for training ML models. When prompting, users directly state in natural language the task they want the pre-trained language model to understand and complete. In contrast, conventional Transformer training first pre-trained models using unlabelled data before fine-tuning them using labelled data for the desired downstream task. A prompt is basically a user-written natural language instruction that the model is supposed to follow. There may be a need for multiple prompts, depending on how difficult the task is that is being trained for. Prompt engineering is the process of selecting the appropriate prompt, or series of prompts, for the required task. Compared to the conventional pre-train & fine-tune method, prompt-based learning has many benefits. The primary benefit is that prompting typically performs quite well with few samples of labelled data.

7.4 Implementation

The proposed implementation can be broadly separated into two major steps:

1. Create prompts for Few-Shot Learning – In this step, a dataset of few-shot sample pairs is created. Each sample pair consists of an outline (o) and corresponding text paragraph (t). The dataset takes the following form:

$$[(o_1, t_1), (o_2, t_2), (o_3, t_3), \ldots, (o_n, t_n)]$$

2. Use the sample pairs as few-shot prompts to generate missing story paragraph for a new outline – The prompt, few-shot samples and the query outline are passed to the model

as input for inference. The model returns the generated story paragraph corresponding to the query outline as prediction.

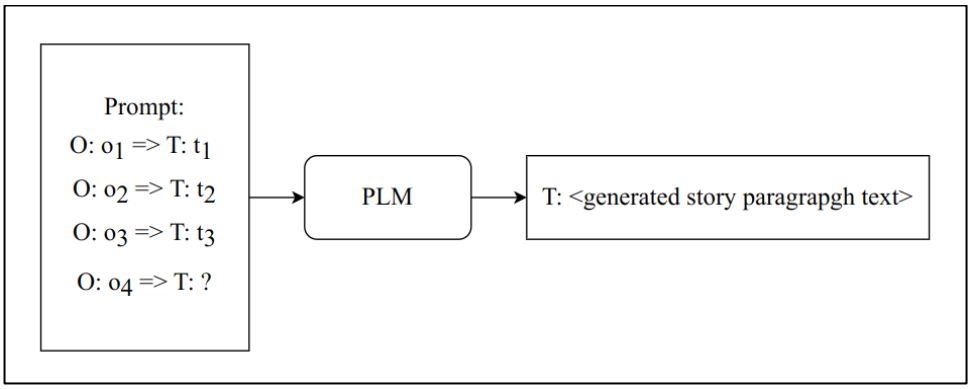

Figure 7.4.1

7.5 Evaluation

The generated stories are to be evaluated using multiple metrics. This work only focuses on evaluation using Automatic Metrics. Human-Evaluation of the generated stories is not within the scope of this work.

The proposed metrics for evaluation are:

- **Perplexity (PPL)** - Similar to (Fang et al., 2021; Jin et al., 2022), PPL is used to compute word-level complexity.
- **DIST/distinct-n** (Li et al., 2015) - DIST measures generation diversity as a ratio of distinct n-grams to all generated n-grams.
- **BLEU** (Papineni et al., 2002) - Measures n-gram overlap between generated text and ground truth.
- **Self-BLEU** (Zhu et al., 2018) - Measures intra-story lexical diversity.
- **ROUGE** (Lin, 2004) - Includes Precision, Recall & F1, where ROUGE Precision has similar interpretation as BLEU score.

This work will be benchmarked against the following baselines:

- Outline-to-Story (**O2S**) (Fang et al., 2021)
- Summarize, Outline and Elaborate (**SOE**) (Sun et al., 2020)
- Prompt Transfer for Text Generation (**PTG**) (Li et al., 2022)

8. Required Resources

8.1 Hardware Requirements

The following hardware requirements must be met for this research work:

- A laptop/desktop computer with internet access capable of browsing, doc-writing and compiling/executing code.
- Access to GPUs to execute CUDA-based deep-learning model training/inference.

8.2 Software Requirements

The following software requirements must be met for this research work:

- Web-browser
- Code IDE
- Python 3.7+
- NVIDIA - CUDA libraries
- Deep Learning libraries such as TensorFlow, PyTorch and HuggingFace
- Other python libraries required for working with data, e.g., Pandas, Numpy, NLTK, etc.

8.3 Dataset Requirements

The following dataset requirements must be met for this research work:

- ROCStories dataset requires a form to be filled and the dataset links are sent via email (ROCStories and the Story Cloze Test, 2022).

9. Research Plan

9.1 Gantt Chart

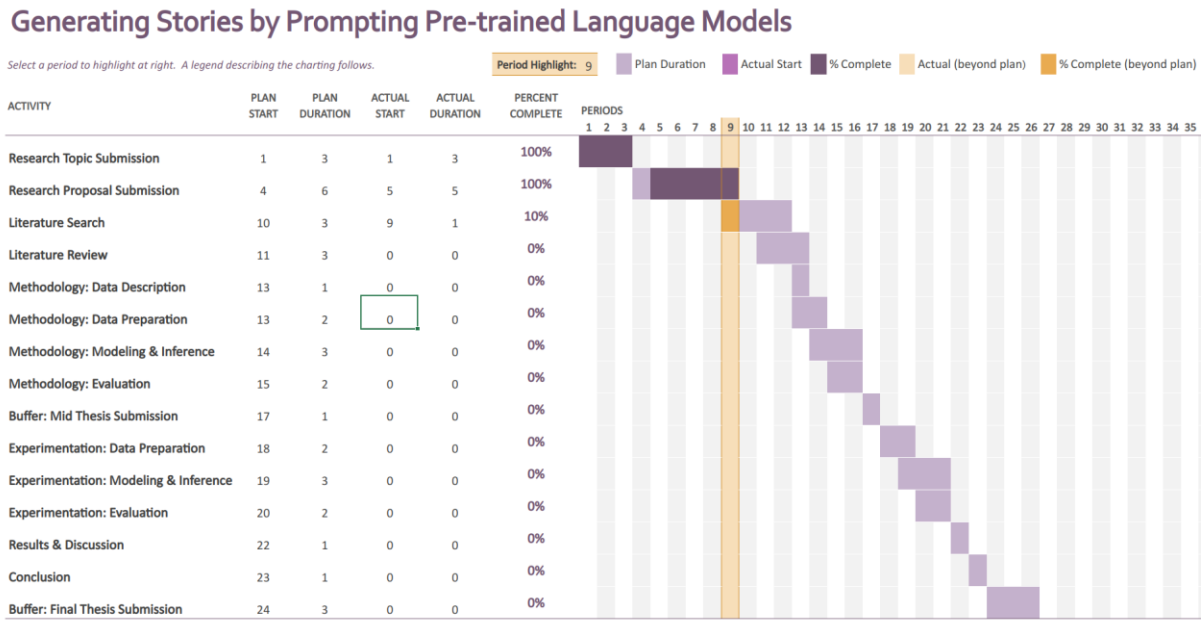

Figure 9.1.1

Note: 1 Period = I Calendar Week

9.2 Risk Mitigation and Contingency Plan

The potential risks to the completion of the thesis work and corresponding contingencies are listed below:

Table 9.2.1

Risk	Contingency
Candidate is unable to perform research work due to health issues or personal problems and it affects timelines.	Plan for buffer time in project management. Inform University/Upgrad administration and ask for extension.
Unavailability of specialized hardware such as GPUs.	Use cloud GPUs.

References

Anon (2022) *ROCStories and the Story Cloze Test*. [online] Available at: https://cs.rochester.edu/nlp/rocstories/ [Accessed 26 Oct. 2022].

Bakhtin, A., Deng, Y., Ott, M., Ranzato, M.' A. and Szlam, A., (2021) Residual Energy-Based Models for Text. *Journal of Machine Learning Research*, [online] 22, pp.1–41. Available at: http://jmlr.org/papers/v22/20-326.html. [Accessed 23 Oct. 2022].

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N. and Turian, J., (2020) Experience Grounds Language. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, [online] pp.8718–8735. Available at: https://arxiv.org/abs/2004.10151v3 [Accessed 23 Oct. 2022].

Bosselut, A., Celikyilmaz, A., He, X., Gao, J., Huang, P. sen and Choi, Y., (2018) Discourse-Aware Neural Rewards for Coherent Text Generation. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, [online] 1, pp.173–184. Available at: https://arxiv.org/abs/1805.03766v1 [Accessed 23 Oct. 2022].

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., Mccandlish, S., Radford, A., Sutskever, I. and Amodei, D., (2020) Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, [online] 33, pp.1877–1901. Available at: https://commoncrawl.org/the-data/ [Accessed 23 Oct. 2022].

Chen, Z., Eavani, H., Chen, W., Liu, Y. and Wang, W.Y., (2019) Few-Shot NLG with Pre-Trained Language Model. [online] pp.183–190. Available at: https://arxiv.org/abs/1904.09521v3 [Accessed 25 Oct. 2022].

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S. and Smith, N.A., (2021) All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, [online] pp.7282–7296. Available at: https://arxiv.org/abs/2107.00061v2 [Accessed 23 Oct. 2022].

Conneau, A. and Lample, G., (2019) Cross-lingual Language Model Pretraining. *Advances in Neural Information Processing Systems*, [online] 32. Available at: https://arxiv.org/abs/1901.07291v1 [Accessed 25 Oct. 2022].

Dathathri, S., Madotto, A., Lan, J., Hung, J., Ai, U., Frank, E., Molino, P., Yosinski, J. and Liu, R., (2019) Plug and Play Language Models: A Simple Approach to Controlled Text Generation. [online] Available at: https://arxiv.org/abs/1912.02164v4 [Accessed 25 Oct. 2022].

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies - Proceedings of the Conference, [online] 1, pp.4171–4186. Available at: https://arxiv.org/abs/1810.04805v2 [Accessed 23 Oct. 2022].

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M. and Hon, H.-W., (2019) Unified Language Model Pre-training for Natural Language Understanding and Generation. *Advances in Neural Information Processing Systems*, [online] 32. Available at: https://github.com/microsoft/unilm. [Accessed 25 Oct. 2022].

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N.A. and Choi, Y., (2021) Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. [online] pp.7250–7274. Available at: https://arxiv.org/abs/2107.01294v3 [Accessed 23 Oct. 2022].

Dziri, N., Madotto, A., Zaiane, O. and Bose, A.J., (2021) Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, [online] pp.2197–2214. Available at: https://arxiv.org/abs/2104.08455v2 [Accessed 23 Oct. 2022].

Fan, A., Lewis, M. and Dauphin, Y., (2018a) Hierarchical Neural Story Generation. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, [online] 1, pp.889–898. Available at: https://arxiv.org/abs/1805.04833v1 [Accessed 23 Oct. 2022].

Fan, A., Lewis, M. and Dauphin, Y., (2018b) Hierarchical Neural Story Generation. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, [online] 1, pp.889–898. Available at: https://aclanthology.org/P18-1082 [Accessed 13 Sep. 2022].

Fan, A., Lewis, M. and Dauphin, Y., (2019) Strategies for Structuring Story Generation. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, [online] pp.2650–2660. Available at: https://arxiv.org/abs/1902.01109v2 [Accessed 23 Oct. 2022].

Fang, L., Li, C., Gao, J., Dong, W. and Chen, C., (2019) Implicit Deep Latent Variable Models for Text Generation. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, [online] pp.3946–3956. Available at: https://arxiv.org/abs/1908.11527v3 [Accessed 25 Oct. 2022].

Fang, L., Zeng, T., Liu, C., Bo, L., Dong, W. and Chen, C., (2021) Outline to Story: Fine-grained Controllable Story Generation from Cascaded Events. [online] Available at: http://arxiv.org/abs/2101.00822 [Accessed 31 Aug. 2022].

Gao, J., Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L. and Shum, H.-Y., (2020a) Robust Conversational AI with Grounded Text Generation. [online] Available at: https://arxiv.org/abs/2009.03457v1 [Accessed 23 Oct. 2022].

Gao, T., Fisch, A. and Chen, D., (2020b) Making Pre-trained Language Models Better Few-shot Learners. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, [online] pp.3816–3830. Available at: https://arxiv.org/abs/2012.15723v2 [Accessed 25 Oct. 2022].

Goldfarb-Tarrant, S., Chakrabarty, T., Weischedel, R. and Peng, N., (2020) Content Planning for Neural Story Generation with Aristotelian Rescoring. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, [online] pp.4319–4338. Available at: https://arxiv.org/abs/2009.09870v2 [Accessed 23 Oct. 2022].

Guan, J., Huang, F., Zhao, Z., Zhu, X. and Huang, M., (2020) A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, [online] 8, pp.93–108. Available at: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00302/43540/A-Knowledge-Enhanced-Pretraining-Model-for [Accessed 23 Oct. 2022].

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R. and Xing, E.P., (2017) *Toward Controlled Generation of Text*. Available at: https://proceedings.mlr.press/v70/hu17e.html [Accessed 25 Oct. 2022].

Hua, X. and Wang, L., (2020) PAIR: Planning and Iterative Refinement in Pre-trained Transformers for Long Text Generation. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, [online] pp.781–793. Available at: https://arxiv.org/abs/2010.02301v1 [Accessed 23 Oct. 2022].

Ji, H. and Huang, M., (2021) DiscoDVT: Generating Long Text with Discourse-Aware Discrete Variational Transformer. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, [online] pp.4208–4224. Available at: https://arxiv.org/abs/2110.05999v1 [Accessed 23 Oct. 2022].

Ji, H., Ke, P., Huang, S., Wei, F., Zhu, X. and Huang, M., (2020) Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, [online] pp.725–736. Available at: https://arxiv.org/abs/2009.11692v1 [Accessed 23 Oct. 2022].

Jin, Y., Kadam, V. and Wanvarie, D., (2022) Plot Writing From Pre-Trained Language Models. *INLG 2022*. [online] Available at: https://arxiv.org/abs/2206.03021 [Accessed 24 Aug. 2022].

Keskar, N.S., Mccann, B., Varshney, L.R., Xiong, C., Socher, R. and Research, S., (2019) CTRL: A Conditional Transformer Language Model for Controllable Generation. [online] Available at: https://arxiv.org/abs/1909.05858v2 [Accessed 23 Oct. 2022].

Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, B., (2015) A Diversity-Promoting Objective Function for Neural Conversation Models. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, [online] pp.110–119. Available at: https://arxiv.org/abs/1510.03055v3 [Accessed 29 Oct. 2022].

Li, J., Tang, T., Nie, J.-Y., Wen, J.-R. and Zhao, W.X., (2022) Learning to Transfer Prompts for Text Generation. *NAACL 2022*, [online] pp.3506–3518. Available at: https://arxiv.org/abs/2205.01543v2 [Accessed 11 Oct. 2022].

Li, J., Tang, T., Zhao, W.X., Wei, Z., Yuan, N.J. and Wen, J.R., (2021) Few-shot Knowledge Graph-to-Text Generation with Pretrained Language Models. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, [online] pp.1558–1568. Available at: https://arxiv.org/abs/2106.01623v1 [Accessed 25 Oct. 2022].

Li, X.L. and Liang, P., (2021) Prefix-Tuning: Optimizing Continuous Prompts for Generation. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, [online] pp.4582–4597. Available at: https://arxiv.org/abs/2101.00190v1 [Accessed 25 Oct. 2022].

Lin, C.-Y., (2004) *ROUGE: A Package for Automatic Evaluation of Summaries*. Available at: https://aclanthology.org/W04-1013 [Accessed 29 Oct. 2022].

Lin, Z. and Riedl, M.O., (2021) Plug-and-Blend: A Framework for Controllable Story Generation with Blended Control Codes. *NAACL | NUSE | 2021*, [online] pp.62–71. Available at: https://aclanthology.org/2021.nuse-1.7 [Accessed 31 Aug. 2022].

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z. and Tang, J., (2021) GPT Understands, Too. [online] Available at: https://arxiv.org/abs/2103.10385v1 [Accessed 25 Oct. 2022].

Mao, H.H., Majumder, B.P., McAuley, J. and Cottrell, G.W., (2019) Improving Neural Story Generation by Targeted Common Sense Grounding. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, [online] pp.5988–5993. Available at: https://aclanthology.org/D19-1615 [Accessed 31 Aug. 2022].

Mihalcea, R. and Tarau, P., (2004) TextRank: Bringing Order into Texts.

Mori, Y., Yamane, H., Shimizu, R. and Harada, T., (2022) Plug-and-Play Controller for Story Completion: A Pilot Study toward Emotion-aware Story Writing Assistance. [online] pp.46–57. Available at: https://aclanthology.org/2022.in2writing-1.6 [Accessed 1 Oct. 2022].

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P. and Allen, J., (2016) A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, [online] pp.839–849. Available at: https://aclanthology.org/N16-1098 [Accessed 13 Sep. 2022].

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J., (2002) Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, [online] pp.311–318. Available at: https://aclanthology.org/P02-1040 [Accessed 29 Oct. 2022].

Pascual, D., Egressy, B., Bolli, F. and Wattenhofer, R., (2020) Directed Beam Search: Plug-and-Play Lexically Constrained Language Generation. [online] Available at: http://arxiv.org/abs/2012.15416 [Accessed 31 Aug. 2022].

Pascual, D., Egressy, B., Meister, C., Cotterell, R. and Wattenhofer, R., (2021) A Plug-and-Play Method for Controlled Text Generation. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, [online] pp.3973–3997. Available at: https://aclanthology.org/2021.findings-emnlp.334 [Accessed 31 Aug. 2022].

Peng, N., Ghazvininejad, M., May, J. and Knight, K., (2018) Towards Controllable Story Generation. [online] pp.43–49. Available at: https://aclanthology.org/W18-1505 [Accessed 25 Oct. 2022].

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., (2019) Language Models are Unsupervised Multitask Learners. [online] Available at: https://github.com/codelucas/newspaper [Accessed 23 Oct. 2022].

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, [online] 21, pp.1–67. Available at: http://jmlr.org/papers/v21/20-074.html. [Accessed 25 Oct. 2022].

Rashkin, H., Celikyilmaz, A., Choi, Y. and Gao, J., (2020) PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, [online] pp.4274–4295. Available at: https://aclanthology.org/2020.emnlp-main.349 [Accessed 31 Aug. 2022].

Rashkin, H., Smith, E.M., Li, M. and Boureau, Y.L., (2018) Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, [online] pp.5370–5381. Available at: https://arxiv.org/abs/1811.00207v5 [Accessed 23 Oct. 2022].

Rose, S., Engel, D., Cramer, N. and Cowley, W., (2010) Automatic Keyword Extraction from Individual Documents. *Text Mining: Applications and Theory*, [online] pp.1–20. Available at: https://onlinelibrary.wiley.com/doi/full/10.1002/9780470689646.ch1 [Accessed 29 Oct. 2022].

le Scao, T. and Rush, A.M., (2021) How Many Data Points is a Prompt Worth? *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, [online] pp.2627–2636. Available at: https://arxiv.org/abs/2103.08493v2 [Accessed 25 Oct. 2022].

See, A., Pappu, A., Saxena, R., Yerukola, A. and Manning, C.D., (2019) Do Massively Pretrained Language Models Make Better Storytellers? *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, [online] pp.843–861. Available at: https://arxiv.org/abs/1909.10705v1 [Accessed 23 Oct. 2022].

Shakeri, H., Neustaedter, C. and DiPaola, S., (2021) SAGA: Collaborative Storytelling with GPT-3; SAGA: Collaborative Storytelling with GPT-3. *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. [online] Available at: https://doi.org/10.1145/3462204.3481771 [Accessed 1 Oct. 2022].

Shen, T., Lei, T., Barzilay, R., Jaakkola, T. and Csail, M., (2017) Style Transfer from Non-Parallel Text by Cross-Alignment. *Advances in Neural Information Processing Systems*, [online] 30. Available at: https://github.com/shentianxiao/language-style-transfer. [Accessed 25 Oct. 2022].

Shin, T., Razeghi, Y., Logan, R.L., Wallace, E. and Singh, S., (2020) AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, [online] pp.4222–4235. Available at: https://arxiv.org/abs/2010.15980v2 [Accessed 25 Oct. 2022].

Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.Y., (2019) MASS: Masked Sequence to Sequence Pre-training for Language Generation. *36th International Conference on Machine Learning,*

ICML 2019, [online] 2019-June, pp.10384–10394. Available at: https://arxiv.org/abs/1905.02450v5 [Accessed 25 Oct. 2022].

Su, Y., Wang, X., Qin, Y., Chan, C.-M., Lin, Y., Wang, H., Wen, K., Liu, Z., Li, P., Li, J., Hou, L., Sun, M. and Zhou, J., (2021) On Transferability of Prompt Tuning for Natural Language Processing. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, [online] pp.3949–3969. Available at: http://arxiv.org/abs/2111.06719 [Accessed 25 Oct. 2022].

Sun, X., Fan, C., Sun, Z., Meng, Y., Wu, F. and Li, J., (2020) Summarize, Outline, and Elaborate: Long-Text Generation via Hierarchical Supervision from Extractive Summaries. [online] Available at: https://arxiv.org/abs/2010.07074v2 [Accessed 23 Oct. 2022].

Tan, B., Yang, Z., Al-Shedivat, M., Xing, E.P. and Hu, Z., (2020) Progressive Generation of Long Text with Pretrained Language Models. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, [online] pp.4313–4324. Available at: https://arxiv.org/abs/2006.15720v2 [Accessed 23 Oct. 2022].

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., (2017) Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

Vu, T., Lester, B., Constant, N., Al-Rfou, R., Cer, D. and Research, G., (2021) SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer. [online] pp.5039–5059. Available at: https://arxiv.org/abs/2110.07904v2 [Accessed 25 Oct. 2022].

Xu, J., Ren, X., Zhang, Y., Zeng, Q., Cai, X. and Sun, X., (2018) A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, [online] pp.4306–4315. Available at: https://arxiv.org/abs/1808.06945v2 [Accessed 23 Oct. 2022].

Xu, P., Patwary, M., Shoeybi, M., Puri, R., Fung, P., Anandkumar, A. and Catanzaro, B., (2020) MEGATRON-CNTRL: Controllable Story Generation with External Knowledge Using Large-Scale Language Models. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, [online] pp.2831–2845. Available at: https://arxiv.org/abs/2010.00840v1 [Accessed 23 Oct. 2022].

Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D. and Yan, R., (2019) Plan-and-Write: Towards Better Automatic Storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, [online] 3301, pp.7378–7385. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/4726 [Accessed 23 Oct. 2022].

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y. and Allen, P.G., (2019) Defending Against Neural Fake News. *Advances in Neural Information Processing Systems*, [online] 32. Available at: https://rowanzellers.com/grover [Accessed 23 Oct. 2022].

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J. and Dolan, B., (2019) DialoGPT: Large-Scale Generative Pre-training for Conversational Response

Generation. [online] pp.270–278. Available at: https://arxiv.org/abs/1911.00536v3 [Accessed 23 Oct. 2022].

Zhao, J.J., Kim, Y., Zhang, K., Rush, A.M. and Lecun, Y., (2018) Adversarially Regularized Autoencoders. [online] pp.5902–5911. Available at: https://proceedings.mlr.press/v80/zhao18b.html [Accessed 25 Oct. 2022].

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J. and Yu, Y., (2018) Texygen: A benchmarking platform for text generation models. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pp.1097–1100.

Ziegler, Z.M., Melas-Kyriazi, L., Gehrmann, S. and Rush, A.M., (2019) Encoder-Agnostic Adaptation for Conditional Language Generation. [online] Available at: https://arxiv.org/abs/1908.06938v2 [Accessed 25 Oct. 2022].

Zou, X., Yin, D., Zhong, Q., Yang, H., Yang, Z. and Tang, J., (2021) Controllable Generation from Pre-trained Language Models via Inverse Prompting. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.2450–2460.