

Beijing Opera Script Generation with GPT-2 as A Baseline

Meizhen Zheng, Xiaodong Shi

School of Informatics, Xiamen University, China

Abstract

Beijing opera script generation with machines is an unprecedented research object as far as we know. It is the start work of this project. In this work, we investigate literatures with respect to script generation. We collect a corpus named *jingju* with which we conduct experiments using GPT-2 to study the ability of script generation about the pre-trained model. After recoding the results of the model, we discuss about the performance and envision our future work. Our code is available on GitHub.¹

1 Introduction

Beijing opera is one of the most precious heritage of Chinese traditional culture. While famous and public-known, its researches in the field of computer science such as script auto-generation are pitifully few, when compared to work in other traditional literature (e.g., Chinese Classical Poetry (Li et al., 2020; Yi et al., 2020; Chen et al., 2019) and Chinese Couplet (Yan et al., 2016; Jiang and Zhou, 2008)). We argue that unlike poetry and couplet which are relatively short and rigid-formatted in general, Beijing opera scripts are long and free-written, making it hard to model.

Given the situation in which we have extremely limited preceding work for direct ref-

erence, and absence of relevant dataset, we intend to divide this project into small steps to accomplish. This paper mainly describes the first two step: constructing a corpus and making a baseline.

First, we manage the dataset task. We download a collection of Beijing opera scripts from the Internet, then process them into specific format to fit the baseline model. We name the dataset as *jingju*.

Second, we prepare a baseline. Considering the text nature of Beijing opera script, we view the task as a text generation task, putting the specific features in Beijing opera for future investigation. A task simplified as a pure text generation problem can be dealt with dozens of models. Among these models, OpenAI’s GPT-2 (Radford et al., 2019) has shown to be capable of generating long fluent text in many areas, such as stories (H. Lee et al., 2020), recipes (H. Lee et al., 2020), patent claims (Lee and Hsiang, 2019), and news (Zellers et al., 2019). The amazing capability of GPT-2 appeals us to employ it as our baseline model. We finetune a pretrained GPT-2 with text sequence from *jingju*, then conduct text experiments to get evaluation metrics for future comparison.

¹https://github.com/zhengmidon/jingju_baseline

Our contributions are as follows:

- We collect a dataset named *jingju* which contains 1215 Beijing opera scripts.
- We construct a baseline with pretrained GPT-2 for further investigation about our project.

2 Related Work

In this section we refer to some related work with respect to the task for future reference.

2.1 Script Generation

Script Generation task defines a problem that one or more agents generate TV play script, theatre play script or other forms of script automatically conditioned on some pre-defined premises. Unlike other text generation task, this topic has not attracted much attention. So we only refer to a bit of work here.

Rosa et al. (2020) and Rosa et al. (2021) employ GPT-2 XL to generate theatre play scripts. Vanilla GPT-2 suffers from several defects so researchers adopt tricks to mitigate problems like repetitiveness and limited context. Especially, they introduce “human in the loop” method as a auxiliary means for a more fluent and coherent generation. While effective to some extent, the proposed agent has not explicitly tackled the problem of coherence and consistence which are common in vanilla GPTs.

Mangal et al. (2019) conducts experiments based on RNNs(LSTM, GRU and Bidirectional RNN) for TV series script generation. They simply use language modelling to generate scripts and inevitably fail to learn

some basic connectivity, making it hard to comprehend the context and plot.

Zhu et al. (2020) propose a model that writes movie scripts conditioned on narratives(i.e., prompts). They treat script generation task as a dialogue generation system and incorporate retrieval-based method to deal with the generation process. To handle the problem of progress tracking, they propose a specific updating mechanism. The most attractive highlight in this work is that a novel cross-attention mechanism between narrative, context and candidate is introduced to model the relation among these three parts for a more coherent plot construction. While this agent is not able to directly generation dialogue in scripts, it is a valuable reference for our project.

2.2 Dialogue Systems

Dialogue Systems (DS) usually structure dialogues in turns from an intelligent agent interacting with human. Existing work generally divided dialogue systems into two types on the basis of their functional positioning: *task-oriented dialogue systems* and *open-domain dialogue systems*.

The task-oriented dialogue system is also called Closed Domain Dialogue System or Goal Driven Dialogue System, which means that the system has clear service goals or service objects, such as querying restaurants, querying bus lines, querying weather, booking tickets, and ordering meal. We didn't refer to this kind of work for its low relevance with our project.

Open-domain dialogue systems display a

more unstructured conversation, as their purpose is to have open-domain dialogues with no specific task to solve. Most of these systems are built to emulate social interactions. They can be further categorized into two groups: *retrieval-based* and *generation-based*. Retrieval-based methods try to find the most reasonable response from a large repository of conversational data, instead of generating a new one (Wu et al., 2017; Zhou et al., 2018; Zhang et al., 2018). Generation-based systems mainly focus on neural generative conversation models. Sun and Li (2021) conducts a survey on neural dialogue generation methods crossing VAE (Kingma and Welling, 2013), Reinforcement Learning and GAN (Goodfellow et al., 2014), et al. We will focus on this field.

2.3 Story Generation

Another domain relevant to our project is story generation. This kind of systems write parts of or the whole story based on some conditions.

Some of the studies focused on *story ending generation* (Guan et al., 2019; Huang et al., 2021), where the story context is given, and the model is asked to select a coherent and consistent story ending. This is similar to the dialogue generation problem mentioned above. Some recent studies also tried to guide the generation of dialogues (Wu et al., 2019; Tang et al., 2019) or stories (Yao et al., 2019) with *keywords* - the next response is asked to include the keywords. This bears some similarities with our study. Another line of work focuses on providing a *higher level*

semantic representation to the networks and conditioning the generation on it. Martin et al. (2018) and Ammanabrolu et al. (2020) use an event-based representation, where an event roughly represents a clause (predicate, subject, direct and indirect object). The model generates the story at the event level and subsequently realizes the individual events to surface sentences. Multiple authors attempt to increase long-term coherence by *hierarchical* story generation. Fan et al. (2018) generate prompt which guide the whole story generation at the next step, with an encoder-decoder architecture. Yao et al. (2019) take a similar approach, using a “storyline” – a list of entities and items to be introduced in the story in the given order. We attach great importance to *hierarchical structure* because script is hierarchical inherently and the multi-step procedure helps a complicated generation process.

3 Method

In this section, we first frame the formulation of our baseline task. Then we introduce the procedure of dataset construction followed by the experiment settings.

3.1 Task Formulation

Our baseline task is to conduct autoregressive language modelling over our constructed dataset. The basic assumption of autoregressive generation is that the probability of a word sequence equals the product of conditional word probability: $P(x_{1:t}|c) = \prod_{i=1}^t P(x_i|x_{0:i-1}, c)$ where c is the given context, and in this work, c can be a story prompt or the beginning of a conversation.

Dataset	jingju
Language	literary Chinese
#Training samples	27,152
#Test samples	3847
#Development samples	2235
#Avg. words per sample in Training	991
#Avg. words per sample in Test	992
#Avg. words per sample in Development	988

Table 1: Detailed information of the filtered jingju dataset.

3.2 Dataset Construction

We download a collection of Beijing opera scripts from website². It contains 1215 scripts which tell stories throughout the whole Chinese history, and is still updating. The scripts are written in literary Chinese with length range from dozens to tens of thousands characters. Each script is composed of introduction of characters, a brief scenario and the main plot divided into scenes which consist of dialogues between roles.

The initial scripts are in pdf format, we transform them to txt format with a open-source tool³. We clean the text by removing the urls and date marks and some irrelevant tokens, retaining only the scenario and the main plot. In order to help the model identify different dialogue turns, we add a specific token [EOS] after every dialogue. And we insert a [EOP] token to differentiate scenario and main plot. To fit the GPT-2 model whose longest input length is 1024, we split the scripts into chunks consisting of roughly 1024 tokens by three-fourths overlap ratio(“roughly” means that the length is not exact 1024 as we want to retain intact sentences).

²<https://scripts.xikao.com/>

³<https://github.com/euske/pdfminer>

Loss Function	CrossEntropy
Batch Size	5
Tokenizer	BERT Tokenizer
Learning Rate	5e-5
Embedding Size	768
Feedforward Size	3072
Hidden Size	768
Activation	Gelu (Hendrycks and Gimpel, 2016)
Head Number	12
Layer Number	12
Dropout	0.1

Table 2: Detailed configurations of the finetuned model.

We randomly sample scripts to form development set and test set, the remains consist the training set. We name the dataset as *jingju*.

Statistics for this dataset are detailed in Table 1

3.3 Experiment Settings

As is analyzed in section 1, we employ GPT-2 as our baseline model. From Du (2019) we find a GPT-2 model pretrained with 1.5 GB ancient Chinese corpora. It is a GPT-2-Small model with input sequence length of 1024. One can refer to the url⁴ for training details. The model is mentioned as *gpt2-chinese-ancient* in the following.

For faster pretrained model construction, we make use of the pretraining framework proposed by Zhao et al. (2019), the Uer-py. It is a framework based on PyTorch (Paszke et al., 2019). The *gpt2-chinese-ancient* model is downloaded from huggingface (Wolf et al., 2020) repository and transformed to Uer-py format which can be processed by the framework. We tokenize training set, development set and test set with BERT (Devlin et al., 2018) tokenizer. We employ cross entropy as

⁴<https://huggingface.co/uer/gpt2-chinese-ancient>

the loss function, finetune the pretrained *gpt2-chinese-ancient* model in a single RTX Titan GPU for 10k steps, and validate the performance of the model in development set for early stopping every 1k steps. Learning rate is set to $5e-5$, batch size is 5. The detailed configurations are listed in Table 2.

4 Results

4.1 Main Results

Following Bai et al. (2021) which also explore the performance of GPT-2 in simple text generation task, we only employ automatic metrics (i.e., F1, Perplexity, BLEU (Papineni et al., 2002), Distinct (Li et al., 2016)) to evaluate the performance of our model on test set. Please refer to Table 3 for exact data. In particular, we truncate the former 100 tokens of every sample in test set as the input context, and the remains as the reference, for the test of BLEUs. Given that autoregressive generation of the model is a time-consuming process, we only test BLEUs and DISTs on 1000 samples in test set.

For intuitive experience of the performance about our baseline model for Beijing opera script generation, we conduct two generation instances on the finetuned GPT-2 given prompts sampled from 《世说新语》 and scenario from script outside the training set. We employ *top-k* and *top-p* sampling for a stable generation. Following Al-Rfou et al. (2019), we use the sliding window method to generate text longer than 1024 tokens. See Appendix A for details.

4.2 Discussion

DIST1 metric of the baseline model suffers a extremely low score, we argue that it is because Chinese has limited character set, and DIST1 penalizes repetitiveness of every character in Chinese case.

We expect the GPT-2 to generate long content that is relevant to the given scenario which describes the story line. The experimental results turn that the model is not capable of managing such complicated task. The features of the dialogues generated by our finetuned GPT-2 share great similarity with that from Rosa et al. (2020). It can produce a text that follows the formal structure and has some level of content coherence. However, the basic attributes of a dramatic situation are missing:

- There is no plot, and the scene is not moving towards a conclusion.
- Having new characters appear randomly in the middle of the scene. And lack of definition about the characters results in lack of personality of them. In fact, they even don't know who they are.
- The content suffers from low logic.
- Lack of long term coherence which is the main drawback of the generation ability about the model.

We argue that GPT-2 is inherently good at imitating the *language style* of the training data and generating neat single sentences. However, because of the simple pretraining target and absence of auxiliary structured knowledge or structured constraints, it is not able

ParaType	F1	PPL	BLEU1	BLEU2	DIST1	DIST2
Baseline	52.05%	7.84	48.69%	31.12%	0.43%	11.28%

Table 3: Test results of different metrics. PPL indicates Perplexity, DIST indicates Distinct metric (Li et al., 2016)

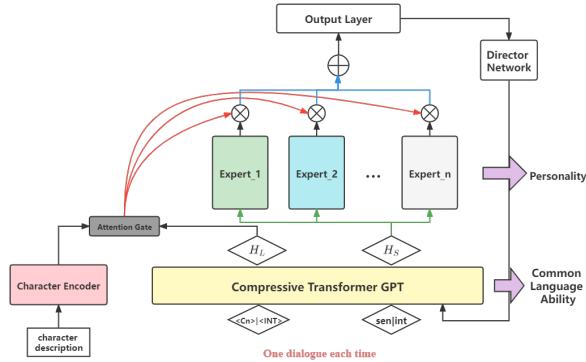


Figure 1: Our preliminary scheme for future work

to learn how to deal with the problem mentioned above. So we will further improve our baseline model in these aspects.

5 Conclusion and Future Work

The project is currently at its official start, with some preliminary work done. In this paper, we refer to some work relevant to our research topic, construct a *jingju* corpus which is used for finetuning a GPT-2 model pretrained with literary Chinese. We baseline the model and test some automatic metrics with it for future comparison.

The analyses in subsection 4.2 indicate the direction of our future work. First, inspired by [Zhu et al. \(2020\)](#), we tend to treat our project as dialogue generation task given the multi-round-dialogue nature of Beijing opera scripts. We will dive into the dialogue system domain for more inspiring methods. Second, consistence and coherence will be

key points to which we attach great importance. We consider incorporating some structured constraints into our model, maybe generating script in hierarchical fashion. Third, noting that character feature is one of the most significant parts of the theatre art(an opera essentially tells *who* do *what* in *somewhere*), we plan to introduce character personality network to model attributes of roles.

We depict our preliminary scheme in Figure 1.

References

- Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166.
- Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7375–7382.
- He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu, and Ming Li. 2021. Semantics of the unwritten: The effect of end of paragraph and sequence tokens on text generation with gpt2. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 148–162.
- Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *IJCAI*, pages 4925–4931.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Zeyao Du. 2019. Gpt2-chinese: Tools for training gpt2 model in chinese language. <https://github.com/Morizeyao/GPT2-Chinese>.

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7375–7382.

He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu, and Ming Li. 2021. Semantics of the unwritten: The effect of end of paragraph and sequence tokens on text generation with gp-t2. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 148–162.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *IJCAI*, pages 4925–4931.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zeyao Du. 2019. Gpt2-chinese: Tools for training gpt2 model in chinese language. <https://github.com/Morizeyao/GPT2-Chinese>.

- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.
- Helena H. Lee, Ke Shu, Palakorn Achananuparp, Philips Kokoh Prasetyo, Yue Liu, Ee-Peng Lim, and Lav R Varshney. 2020. RecipeGPT: Generative pre-training based cooking recipe generation and evaluation system. In *Companion Proceedings of the Web Conference 2020*, pages 181–184.
- Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.
- Qingbao Huang, Linzhang Mo, Pijian Li, Yi Cai, Qingguang Liu, Jielong Wei, Qing Li, and Ho-fung Leung. 2021. Story ending generation with multi-level graph convolutional networks over dependency trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13073–13081.
- Long Jiang and Ming Zhou. 2008. Generating chinese couplets using a statistical mt approach. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 377–384.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Jieh-Sheng Lee and Jieh Hsiang. 2019. Patent claim generation by fine-tuning openai gpt-2. *arXiv preprint arXiv:1907.02052*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.
- Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. 2020. Rigid formats controlled text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 742–751.
- Sanidhya Mangal, Poorva Joshi, and Rahul Modak. 2019. Lstm vs. gru vs. bidirectional rnn for script generation. *arXiv preprint arXiv:1908.04332*.
- Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Rudolf Rosa, Ondřej Dušek, Tom Kocmi, David Mareček, Tomáš Musil, Patrícia Schmidtová, Dominik Jurko, Ondřej Bojar, Daniel Hrbek, David Košťák, et al. 2020. Theaitre: Artificial intelligence to write a theatre play. *arXiv preprint arXiv:2006.14668*.
- Rudolf Rosa, Tomáš Musil, Ondřej Dušek, Dominik Jurko, Patrícia Schmidtová, David Mareček, Ondřej Bojar, Tom Kocmi, Daniel Hrbek, David Košťák, et al. 2021. Theaitre 1.0: Interactive generation of theatre play scripts. *arXiv preprint arXiv:2102.08892*.
- Bin Sun and Kan Li. 2021. Neural dialogue generation methods in open domain: A survey. *Natural Language Processing Research*, 1(3-4):56–70.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in*

- Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.
- Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505.
- Rui Yan, Cheng-Te Li, Xiaohua Hu, and Ming Zhang. 2016. Chinese couplet generation with neural network structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2347–2357.
- Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.
- Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9450–9457.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9051–9062.
- Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752.
- Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.
- Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127.
- Yutao Zhu, Ruihua Song, Zhicheng Dou, Jian-Yun Nie, and Jin Zhou. 2020. Scriptwriter: Narrative-guided script generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8647–8657.

A Case Study

The prompt of the first example Case A.1 is selected from a script that the model has not seen, the second A.2 is from 《世说新语》 which is a fiction written in Song dynasty. We follow Al-Rfou et al. (2019), using sliding window strategy when generate dialogues longer than 1024 tokens.

A.1 Case 1

Input prompt:

大宋年间，丞相包拯审理狸猫换太子案，回京途中路经河北太平县程南庄地带，在轿中忽做一梦，梦见一冤魂前来告状，冤魂哭诉他本是两年前去幽州贩马的客商，路过程南庄夜宿灶王庙，遇到庄上买芽糖为生的恶徒龚金笑，那恶徒见财起意，为劫取他三匹枣红赤兔马，趁其不备用切刀将其杀害致死。为防被人发现，那恶徒便将其尸骨埋葬在他家后院，乞求包拯与其伸冤。

Output of the baseline model:

【第一场】

（张旺上。）

张旺（引子）三载离家，贩卖绸缎为生，头戴乌云鬓，身穿大红白，四季衣裳换。

（念）生意兴隆通四海，财源茂盛达三江。人前称赞我恩德，早与人家庆团圆。

（白）下官张旺。幼读诗书，幼习兵戈，蒙丞相恩养成性，官居总戎之职。今当早年，我只得在此侍候。

（张旺下。）

【第二场】

（二太监引宋仁宗同上。）

宋仁宗（引子）海晏河清，四海升平。

（念）万民乐业享安宁，五谷丰登保太平。孤王可称有道君，只恨陈州起战争。

（白）孤，宋仁宗在位。只因郭槐，奉旨监斩陈州百姓，还不见到来。内侍，闪放银安。

（宋仁宗下。）

【第三场】

张旺（内白）马来。

（张旺上。）

张旺（念）奉旨监斩贼臣首，杀死二老在朝门。

（白）我，张旺。奉了万岁之命，监斩陈州百姓，也不知被哪家御史台前参奏，查出奴婢名姓，搜出马匹，我在包拯面前连连说词的那贼，就将女儿抢去。那包拯不容我开口，反命钦差大臣审问。今日闲暇无事，不免在此伺候。

（二幕外。陈士美持马上。）

陈士美（念）只为陈州事，常到包拯门。

（白）陈士美。奉了万岁之命，监斩陈州百姓。来，请王爷。

（王朝上。）

王朝（念）耳旁又听得人声喊，想必包拯到来临。

（白）王爷在上，陈士美参见。

宋仁宗（白）王相平身。

陈士美（白）臣，王朝见驾，吾皇万岁。

王朝（白）万万岁。

宋仁宗（白）陈士美平身。

陈士美（白）为臣犯罪，理当见驾，吾皇万岁。

宋仁宗（白）你在此做甚？

陈士美（白）我在此包拯处管告。

宋仁宗（白）但不知你行走如何？

陈士美（白）臣奉命监斩陈州百姓不死。

宋仁宗（白）来，一同开刀。

陈士美（白）有劳公公。回府开看。

（陈士美开刀。）

陈士美（白）哦哦！

王朝（白）明公请息雷霆之怒。

陈士美（白）不忠不孝之人，何必动怒！

王朝（白）明公请息雷霆之怒。小人若是开口，先斩后奏。

刘后（白）我想那包拯乃是铁面无私之人，你莫非是张千岁当年的冤家吗？

宋仁宗（白）不敢。卿家你休要如此，待孤从长计议。

刘后（白）小妾启奏万岁，包拯乃当朝太师，寡人亦非他人可比。

宋仁宗（白）你休得如此。你前番进得相府，见了你家夫人，奏知寡人，认你为义子，也就是了。

刘后（白）但凭万岁。

宋仁宗（白）退班！

（四太监、大太监、宋仁宗同下。）

刘后（二黄原板）想当年在相府有恩有德，因此上一家人定然欢悦。

（刘后下。）

【第四场】

（二刽子手同上，同跑下。王朝上。）

刘后（二黄原板）叫刽子手快抬头看，只见那刽子手面带雕。上面坐的二将官，文武大小将官员个个头戴，一品的忠良是哪个？孤家何不将他赶，赶出了包拯世外头。

（四刽子手、二太监同上，同搀刘后下。王朝下，四刽子手同下。李文忠上。）

李文忠（二黄原板）手掌棍朝下打，耳听得御史奏一本。将身上了龙圈椅，不见他怎把路趲。

（二刽子手同上，同跑下。李文忠下，四刽子手同下。）

【第五场】

（二刽子手、二太监押包拯同上，包拯不见。四刽子手押王朝、马汉、张龙、赵虎、董超、薛霸同上，同进城，同下。宋仁宗上。）

包拯（二黄原板）见此景不由人心内焦躁，倒叫我心胆寒。

（白）且住。方才万岁之言，包拯是当年的冤家。但不知他那里出了什么妖魔？

宋仁宗（白）此人有何变化？

包拯（白）万岁不要错疑。此人身为当朝一品，忠心为国。万岁命臣监斩，怎奈万岁不允。如今万岁命臣监斩，臣前来与他二人辩理。

宋仁宗（白）你二人此去，定然是打从水门经过。

包拯（白）想必是斩草除根。

宋仁宗（白）但不知斩草除根？

包拯（白）万岁不可。

宋仁宗（白）啊？

包拯（白）万岁不必多疑，快快出宫。

（宋仁宗出宫。）

宋仁宗（白）且住。方才万岁言说：斩草除根，乃是一件。

（宋仁宗哭，包拯哭。）

宋仁宗（哭）哎呀！包拯你哭死我了！

（二黄摇板）这才是为国忧民，包拯（二黄摇板）这也是为国忧民。

宋仁宗（二黄摇板）包拯离宫院三千里，去到五殿救娘亲。辞别万岁出宫门，我是再造伴驾臣。

（包拯下。）

宋仁宗（二黄摇板）好一个忠良臣，忠心保国保朝廷。内臣摆驾进宫廷，不知有何议论？

（宋仁宗下。）

...

A.2 Case 2

Input Prompt:

济阳祝村有祝翁者，年五十余病卒，家人入室理纓经，忽闻翁呼甚急。群奔集灵寝，则见翁已复活，群喜慰问。翁但谓媼曰：“我适去，拚不复还。行数里，转思抛汝一副老皮骨在儿辈手，寒热仰人，亦无复生趣，不如从我去。故复归，欲偕尔同行也。”咸以其新苏妄语，殊未深信。翁又言之。媼云：“如此亦善。但方生，如何使死？”翁挥之曰：“是不难。家中俗务，可速料理。”媼笑不去，翁又促之。乃出户外，延数刻而入，绉之曰：“处置妥妥矣。”翁命速妆，媼不去，翁催益急。媼不忍拂其意，遂裙妆以出，媳女皆匿笑。翁移首于枕，手拍令卧。媼曰：“子女皆在，双双挺卧，是何景象？”翁捶床曰：“并死有何可笑！”子女见翁躁急，共劝媼姑从其言。媼如言，并枕僵卧，家人又共笑之。俄时媼笑容忽敛，又渐而两眸俱合，久之无声，俨如睡去。众始近视，则肤已冰而鼻无息矣。视翁亦然，始共惊怛。

Output of the baseline model:

【第一场】

（祝英台上。）

祝英台（引子）小子业已多沦落，从早来依贫是我作新郎。

（念）为避嫌，小字祝英台。父嫌乡里尊，父名叫英台。才长已成人，又深学圣贤。全凭一纸书，读过三坟五典。

（白）我乃前辈先生。闻听父亲，欲入京赴考，不意天降雷霆，是我命人传令，将祝英台发配云南。怎奈路途遥远，行走不便，是我将他改与英台，且命他在外寻找。看前面已是云南，他若至此，只恐他难以飞过。不免把我唤出来，劝他归家相见便了。我在此等候，我去叫媒婆去便了。

（祝英台下。媒婆上。）

媒婆（念）忽听媒婆唤，急忙到跟前。

（白）媒婆参见太老爷。

（诸位头目同上。）

诸位头目（同白）叩见太老爷。

祝英台（白）你等在哪里？媒婆。

诸位头目（同白）在。

祝英台（白）你等可晓得祝英台在哪里？

诸位头目（同白）这个……

祝英台（白）姐姐请坐。

诸位头目（同白）告坐。此处甚是僻雅，前面就是云南，不知什么地方，无有什么地方。

祝英台（白）唉！人言道此处是要买房屋，今日看来却是荒僻小乡，怎生到此？

诸位头目（同白）此处非要买房屋。

祝英台（白）何人言道，听她讲些什么？

诸位头目（同白）就是二府衙门。

祝英台（白）啊爹爹，休得要嚷，一同前往。

（〔牌子〕。众人同下。）

【第二场】

（董永上，书童随上。）

董永（引子）爱习拳棒，爱习拳棒，自成霸王。

（念）一步一拐靠山岗，两膀力似浪里。每日拿拳打，惯使双铜。

（白）在下董永。爹爹董永。大比之年上京，一去数载，杳无音信回来。是我在万岁驾前进京，在这涿州一带访友，不觉三载，倒也十分潇洒。今日天气清和，不免前去走遭也。

（西皮摇板）二堂领了太爷命，涿州城内访英雄。

（白）来此已是涿州地方，不免在此安排一番便了。

（西皮摇板）忽听帐内传唤定，想必来此访英雄。

（白）看天上已是王府，不免驾云前往。

（西皮摇板）迈开大步朝前进，又只见一位在路旁。

（白）里面有人么？

（院子上。）

院子（白）是哪一位？

书童（白）原来是大相公。你进去问问那里的买卖，可有大小的买卖？

院子（白）有。

书童（白）大概都有。

院子（白）请少待。启禀相爷、太老夫人。

（媒婆上。）

媒婆（白）参见相爷、太老夫人！

董永（白）罢了。到此何事？

媒婆（白）前次与张小姐婚姻。

董永（白）不用办啦，到了那里，你我告辞了。

（）

家院（白）送太老爷，请他进来。

（院子下。董永、青袍同下。）

董永（白）媒婆。

媒婆（白）在。

董永（白）张小姐有礼了。

（张小姐上。）

张小姐（白）大人哪，敢莫是为媒婆无有？

董永（白）岂敢。

张小姐（白）今日是三六九日，与张小姐成亲的。

张小姐（白）少备一筵，与大人洗尘。

（张小姐下。）

董永（白）我请张小姐梳洗一回如何？

张小姐（白）奴家也有一陪，今日万岁可要来接你们。

（张小姐上。）

张小姐（白）是。

张小姐（白）姐夫安歇了罢。

张小姐（白）这也不瞒你说，这头一排，有一子名叫柳湘莲，现在睡着了。不非同她梳洗。

（董永下。）

张小姐（白）啊，夫人，这几日未起床，怎生得了安睡呢？

（张永上。）

张永（白）这一排，乃是柳湘莲，昨日托去，小生幸会。不免唤她出来。

（张小姐上。）

张永（白）啊安人，夫人可唤出相公？

董永（白）有劳你了。

张永（白）哼哼，什么事情？

董永（白）今日天气清和，去一小楼，你看天气如何？

张永（白）现在睡着了。何不唤她出来？

张小姐（白）快去，快去。

张永（白）是。

董永（白）张相公，快去，快去。

...