

---

# Controlled Cue Generation for Play Scripts

---

Alara Dirik<sup>†</sup> Hilal Donmez<sup>†</sup> Pinar Yanardag

Boğaziçi University  
Istanbul, Turkey

{alara.dirik, hilal.donmez}@boun.edu.tr  
yanardag.pinar@gmail.com

## Abstract

In this paper, we use a large-scale play scripts dataset to propose the novel task of theatrical cue generation from dialogues. Using over one million lines of dialogue and cues, we approach the problem of cue generation as a controlled text generation task, and show how cues can be used to enhance the impact of dialogue using a language model conditioned on a dialogue/cue discriminator. In addition, we explore the use of topic keywords and emotions for controlled text generation. Extensive quantitative and qualitative experiments show that language models can be successfully used to generate plausible and attribute-controlled texts in highly specialised domains such as play scripts.

## 1 Introduction

Script generation for theater plays involves the automatic generation of a sequence of lines of dialogue and cues that are coherent as a whole. While story and plot generation are relatively popular tasks, play and movie script generation remains a largely unexplored problem. In this paper, we focus on the generation of theatrical cues from character dialogue lines. A theatrical cue can be described as an informative text that is not spoken dialogue. It can be a trigger for an action, an informative description of the stage, thoughts of the characters or body language intended to amplify the effect of the play. Cues are highly variable in context and can range from sound effects, lighting changes, the movement of characters on stage, moods, thoughts, and reactions via silent gestures. The following example illustrates how a cue is used to direct a character’s action on stage and add to their spoken lines.

**JOHN: I don’t know what to do anymore.  
(JOHN turns around and leaves.)**

In addition to describing the actions of the characters, cues also describe the interaction between them, such as the following example:

**LIZZIE: How do you...? (Putting things together:) No . . .**  
**POYDRAS: But you also have her eyes.**  
**LIZZIE: (Weeps. Realizes she is looking at her father. Takes a moment.)**

Theatrical cues play an important role in screenplays by bridging the gap between the audience and the actors. They can set the tone of the scene and conversations and bring otherwise mundane conversations to life. Therefore, cue generation is a useful and valuable tool for playwrights to explore different creative avenues and inspire actors and actresses to present their performances with great impact. Another common use case for theatrical cues is to modernise and reinterpret old plays

---

<sup>†</sup>Equal contribution. Author ordering determined by a coin flip.

without changing the dialogue. In our work, we thoroughly investigate this use case by generating plausible cues based on the original dialogue lines. To this end, we have collected over 1500 play scripts with various topics, containing a total of 775,000 lines of dialogue and over 277,000 cues. To the best of our knowledge, we are the first to propose the novel task of generating cues from dialogues in plays.

In this work, we introduce a new task and use large-scale transformer-based language models trained on large text corpus for controlled text generation. Controlling the attributes of the generated text, such as specific topics or sentiments, remains difficult without fine-tuning the models for each attribute separately. To address this issue, we explore cue generation using the preceding dialogue and propose a cue/dialogue discriminator using the PPLM framework proposed by Dathathri et al. [2020]. We also explore other extensions such as emotion-based and topic-based text generation.

## 2 Related Work

### 2.1 Text generation

Text generation is a very popular NLP task where deep neural networks are widely used, with sequence-to-sequence (seq2seq) (see Sutskever et al. [2014a]) with attention (see Luong et al. [2015]) among the most popular models. Generative adversarial networks (GAN) and autoencoders (see Wang and Wan [2018], Hu et al. [2017b]) have also been used to generate text conditioned on specific attributes. These works focus on training generative models and variational autoencoders for style transfer, which rely on learning disentangled latent representations for style and content.

Most of the work on text generation in recent years has been based on the transformer architecture (see Vaswani et al. [2017], Çelikyilmaz et al. [2020], Hu et al. [2017a], Keskar et al. [2019]), which has enabled training large-scale language models (LMs) on very large datasets and significantly improved the state-of-the-art in natural language processing, as Radford [2018] shows. BERT by Devlin et al. [2018] and GPT-2 by Radford et al. [2019] are among the most successful transformer-based language models. Recent studies have used BERT for conditional text generation, employing a large pre-trained language model to generate text conditioned on intent labels (see Xia et al. [2020]). Similarly, Sheng et al. [2020], Prabhume et al. [2020], Ziegler et al. [2019] have conducted studies on using GPT-2 to generate text with controlled attributes and biases. However, these approaches are often not useful in practice as they require the model to be fine-tuned for each specific attribute separately. In our work, we focus on plug-and-play approaches and generate text by steering pre-trained language models towards acquiring the target attributes.

### 2.2 Story Generation

Previous research in story generation such as Clark et al. [2018] mostly focuses on using recurrent neural networks (RNNs) and long short term memory units (LSTMs) for text generation. However, RNNs have difficulties in generating longer and coherent texts (see Bahdanau et al. [2014], Sutskever et al. [2014b], Cho et al. [2014]), hence other works such as Martin et al. [2018a] aim to provide different semantic representations for story generation.

Martin et al. [2018a] proposed dividing the automated story generation task into two subtasks: successive generation of events (event2event) and generation of human-readable sentences from events (event2sentence). The event2event model generates successive events by extracting semantic information from each sentence and the event2sentence model translates the generated events into human-readable sentences. Controllable story generation (see Peng et al. [2018]) is another text generation method that uses an analyzer consisting of supervised classifiers and rule-based keyword extractors to extract control factors from story corpus and a generator that generates stories with an RNN conditioned on the control factors. While this approach can be used to generate stories that reflect the user’s intent, a separate model needs to be trained for each new intent or control factor.

Interactive story generation is another research area where various machine learning methods have been proposed (see Riedl and Bulitko [2013]). Interactive story generation enables users to influence or direct stories with their inputs. Brahman et al. [2020] focused on the task of interactive story generation, where the user provides mid-level sentence abstractions in the form of cue phrases to the model during the generation process. Akoury et al. [2020] proposed another story generation system

called STORIUM, where human authors query a model for suggested story continuations and edit them.

## 2.3 Dialogue Systems

The rise of deep learning based Natural Language Understanding (NLU) and Natural Language Generation (NLG) methods has significantly improved the performance of dialogue systems. Dialogue systems typically consist of two modules: an NLU module to extract information from user queries and an NLG module to produce relevant responses and start new dialogues. Since dialogue generation directly depends on the performance of the NLU approach used, it is critical to understand the user intent correctly. Vanzo et al. [2019] tried to solve this problem by proposing a hierarchical multitask NLU architecture that creates a domain-independent and rich semantic representation of the user input. This approach aims to encode the structure of the user input along with the actions and arguments it contains via a self-attention mechanism, seq2seq BiLSTM encoders, and CRF tagging layers. Once the user intent is extracted, a conditional text generation method such as a conditional variational autoencoder (see d’Ascoli et al. [2020]) can be used to generate user-intent dependent responses.

## 2.4 Play Script Generation

The vast majority of previous work on creative text generation focuses on song lyrics generation, story generation (see Luo et al. [2019], Jain et al. [2017]), and movie plot and script generation (see Zhu et al. [2020], Martin et al. [2018b], Mangal et al. [2019]), while theater play script generation is explored to a much lesser extent. HTGAA [2017] trained a character-level RNN model on theater play scripts to generate entire plays and stage directions. However, previous work on creative text generation mainly investigates how to generate coherent, reasonable, and diverse stories and scripts. Since creating labeled datasets with the desired attributes is time-consuming and labor-intensive, this work limits the controllability of the generated texts to coarse-grained sentiments (e.g. positive, negative intent). Hence, fine-grained controllable play script generation remains an unexplored topic to the best of our knowledge.

More recently, Rosa et al. [2020] proposed THEaiTRE, a mixed framework that consists of generative language models and hierarchical generation approaches that use text summarization and machine translation methods. THEaiTRE finetunes a pre-trained GPT-2 model Radford et al. [2019] on a small dataset of formatted theater and movie scripts in English and Czech. Moreover, this work proposes to generate a new training dataset by cross-translating between Czech and English to overcome the limited amount of training data. However, it is not possible to evaluate the performance of this approach as the dataset, experimental results and generated play scripts have not been released.

## 3 Dataset

We have collected 1511 English-language play scripts with over 775,000 lines of dialogue and over 277,000 cues on a variety of themes including *Comedy*, *Romance*, *Satire*, and *Greek*. The collected play scripts are scraped from the Playscripts website<sup>3</sup> and usually include the title of the play, production notes, background information on the characters, and the play itself.

A play script is a highly structured text consisting of one or more acts defined by elements such as rising action, climax, and resolution. Each act consists of six or more scenes, with each scene containing conversations between 2-4 characters. While acts represent a broader storyline of interrelated events, a scene usually represents actions that take place in one place and time, and are delineated from the next scene by a curtain, a blackout, or a brief emptying of the stage. Therefore, conversations within a scene are often separate from the preceding scenes and take place between different characters.

Each scene in a play script consists of lines of dialogue and cues. In all play scripts, dialogue lines start with capitalized character names and cue lines are placed in parentheses. In our work, we pre-process raw scripts by eliminating pages that do not contain at least one line of dialogue and one line of cues. Cues are not meant to be spoken aloud by characters, and their lengths are highly variable. They contain stage directions, stage descriptions, and character descriptions that are

---

<sup>3</sup><https://www.playscripts.com>

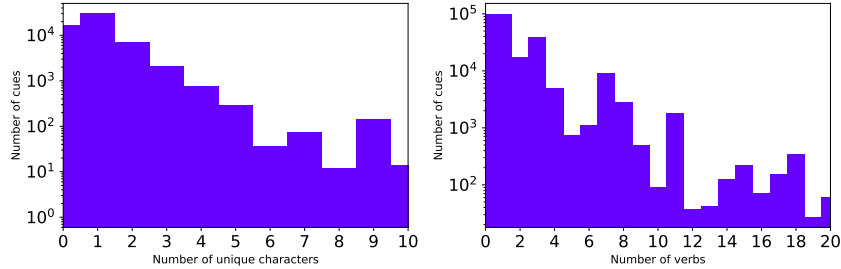


Figure 1: Histogram of number of unique character names (left). Histogram of number of verbs in cues (right).

essential to understanding the scenes, as well as the mood, feelings, and thoughts of the characters conveyed through silent expressions. Cues are valuable tools for actors to communicate with the audience and convey spoken lines/dialogue in a myriad of different ways. In addition, cues are often used to modernise and/or reinterpret plays without changing the dialogue. For example, a cold greeting as opposed to a friendly greeting can say a lot about the relationship between two characters. In addition to indicating the feelings of the characters, cues can also be stage directions such as:

**(Silence as ROLAND exits stage left.)**  
**(LOWELL looks toward the stage right door.)**  
**(GRAHAM runs into the bathroom, stage right. He begins to vomit loudly.**  
**The knocking becomes even more persistent.)**

A manual review of the dataset revealed that stage directions and scene changes make up a small portion of the dataset. To distinguish stage directions and scene changes from the rest of the cues, we counted the number of cues containing the word *stage* and found that 11K out of 227K cues contain the keyword *stage*. The number of character names in the cues varies widely. As shown on the left in Figure 1, some cues contain no character names, while some cues contain up to 10 characters. Cues can also describe actions that characters are supposed to perform (e.g. "Suddenly jumps up from the chair"). To analyze the categories of these actions, we examined the number of verbs that appear in the cues. The right side of Figure 1 shows that some cues contain no action, while some of them can have up to 20 actions.

## 4 Methodology

Plug and Play Language Models (PPLM) aim to leverage large pre-trained language models (LM) to generate attribute controlled text without fine-tuning or re-training the models. In the context of our work, controllable generation refers to modeling the conditional likelihood of generated text  $p(x|a)$ , where  $a$  denotes desired controllable attribute(s) such as emotion, topic, sentence type/intent and  $x$  is the generated sample. PPLM plugs an attribute model  $p(a|x)$  together with a base generative model  $p(x)$  (GPT-2) and sample from the resulting conditional likelihood  $p(x|a) \propto p(a|x)p(x)$ . Therefore, it effectively creates a conditional generative model on the fly from any given attribute model, where the attribute models are either in the form of a bag-of-words (BoW) or a discriminator with a single learned layer, without any further training of the underlying base LM.

The PPLM method uses GPT-2 medium as the base LM, which is a left-to-right autoregressive model that generates one token at a time, using the preceding text as input. Given a sequence of tokens or preceding text  $\{x_0, \dots, x_{n-1}\}$ , transformer based LMs compute the unconditional probability of the resulting sequence  $p(X)$  for all succeeding token candidates:

$$p(X) = \prod_{i=1}^n p(x_i | x_0, \dots, x_{i-1}) \quad (1)$$

Moreover, the GPT-2 architecture uses the hidden representation  $H_t$  to generate  $x_{t+1}$ , given  $x_t$ . In order to steer the output of the LM, PPLM shifts the hidden representations  $H_t$  towards the sum of two gradients at each generation step  $t$ : towards the higher log-likelihood of attribute  $a$  under the

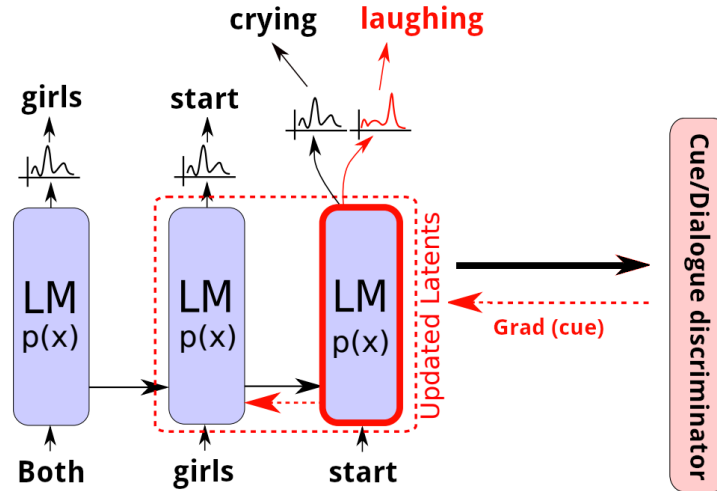


Figure 2: An illustration of the PPLM approach with cue/dialogue discriminator. Figure is modified from Dathathri et al. [2020].

conditional attribute model  $p(a|x)$ , and towards the higher log-likelihood of the base LM  $p(x)$ . Thus, the shifted hidden representation ( $H_t + \Delta H_t$ ) leads to a distribution of generated text that is more likely to contain the selected attribute(s). As in the original PPLM experiments, we initialize  $\Delta H_t$  to zero and update it with gradients from the attribute model that measures the closeness between the generated text and the desired attribute such as a topic, emotion, intent.

Furthermore,  $\Delta H_t$  is updated to minimize the KL divergence between the output distribution of the modified and unmodified language models to ensure fluency. In addition to minimizing KL divergence, post-norm fusion is performed similarly to Stahlberg et al. [2018] to bind the generated text to the unconditional  $p(x)$  LM distribution.

We note that the baseline PPLM framework uses only seven manually generated lists of topic words and a sentiment discriminator trained on the IMDB movie reviews dataset (see Maas et al. [2011]), which is insufficient for our task. Therefore, we use PPLM as our base framework and train a cue/dialogue sentence type discriminator to condition the generation towards cues (see Figure 2). In addition to the cue/dialogue sentence type discriminator, we introduce and experiment with two other attribute models: an automated topic modeling module and an external multi-label emotion classifier, Deep-Moji (see Felbo et al. [2017]), for controlled text generation. While the dialogue/cue discriminator and the topic-based approach aim to generate appropriate cues, the emotion classifier is used to steer the generated text towards the emotion label of the input text. We describe the details of the three attribute models we use in Sections 4.1, 4.2 and 4.3.

#### 4.1 Controlled Generation using Cue/Dialogue Discriminator

We train a binary cue/dialogue discriminator using 10% of our dataset, where the input sentences  $x$  are tagged with their corresponding labels  $y$ . The discriminator consists of a single-layer classifier that predicts the target label. Based on the sentence type given as input by the user and the classifier prediction, PPLM shifts the activations towards the higher log-likelihood of either the dialogue lines or cues as specified by the user.

#### 4.2 Controlled Generation using LDA

Unlike PPLM, where the lists of topic keywords are created manually, we create word lists by automatically extracting topics using Latent Dirichlet Allocation (LDA) (see Blei et al. [2003]). To this end, we create a cue corpus and model it as a distribution of 10 topics. We use the trained LDA model to extract the topic of the input sentences (dialogue lines) and dynamically steer the inference to generate cues that have the same topic as the input sentence. A target topic selected by the user is then used to steer the language generation process to maximize the log-likelihood of the extracted target topic keywords.

Table 1: Model output samples for GPT-2+FT and PPLM+CueDisc methods.

Input	GPT-2+FT	PPLM+CueDisc
OWEN. Yeah. He’s awesome.	MEG. Yeah. He is. He has the biggest smile on earth... OWEN. Oh.	WADSWORTH. You think he likes me, too? OWEN. (Laughs, but then gets serious: )
VAL. He’s been looking at me.	STEWART. Who is he? VAL. I don’t know, never seen him.	(VAL stares back at JOE, who looks confused.) VAL. I said, “What’s his problem?”
JACKIE. Okay, I can do this. Thanks.	RENEE. You can do this? You can do this!	(JACKIE starts to walk away. The other girls, all laughing, stop her.)
CAL. My mother is dead.	MADELINE. She died of a drug overdose. CAL. That’s a lie.	(She pulls back the sheet.) I don’t want to know how.
STEWART. Yes! Yes, it is. This is it!	The moment. You know what I mean. We have to get out of here right now!	MISS BLAINE. It’s not. STEWART. (Suspiciously:) What do you mean?

### 4.3 Controlled Generation with Emotions

Since plays contain a wide range of emotions, and not just positive or negative sentiments, we train an emotion classifier using DeepMoji, a sentiment model that predicts the emoji label of input sentences (see Felbo et al. [2017]). We use DeepMoji to predict the emojis corresponding to the given lines, and then map the predicted emojis to a subset of emotions from Plutchik’s Wheel of Emotions (see Plutchik [1980]). We then use the input sentences (dialogue lines) and their corresponding emotion labels to train an emotion classifier. The trained classifier is used to steer the generation towards the target emotion label and does not necessarily generate cues.

### 4.4 Experimental Setup

We compare the PPLM-based extensions with fine-tuned GPT-2 and Infilling by Language Modeling (ILM) Donahue et al. [2020] baselines. For the PPLM experiments, we use the GPT-2 345M model fine-tuned on 80% of our dataset as our base generative model  $p(x)$ . While the GPT-2 model is not fine-tuned in the original work, the structure and rigid syntax of play scripts require fine-tuning the model to generate plausible dialogues and cues. We use 10% of the dataset to perform steered inference and test the PPLM approaches, and the remaining 10% to train the conditional attribute models  $p(a|x)$ : a binary cue/dialogue classifier, a multi-label emotion classifier, and for topic modeling via LDA to be used in the PPLM experiments. We also perform a simple preprocessing step to insert a white space between the punctuation and the alphanumeric characters. For the ILM experiments, we first divide our dataset into training, validation and test sets in a ratio of 80-10-10 and create infilling examples. To do this, we divide the dataset into successive triples of lines, where the lines can be either dialogues or cues in any order. We also randomly mask paragraphs, sentences, n-grams, and words with a masking probability of 3% each, resulting in a marginal token masking rate of 15%. For fine-tuning, we insert a bos token  $\langle BOS \rangle$  at the beginning of each scene and an eos token  $\langle EOS \rangle$  at the end of each scene to mark the beginning and end of different conversations. We filter the training and test datasets to only include consecutive dialogue-cue-dialogue triplets and use the start and end dialogue lines as input during inference.

- **GPT-2+ FT** : Given a line of dialogue as input, we use a GPT-2 model fine-tuned on our dataset to generate text.
- **ILM**: ILM enables LMs to infill variable-length spans using both preceding and subsequent text. We follow the same approach proposed in the ILM paper and fine-tune the GPT-2 small model on successive line triples following the order dialogue-cue-dialogue. The second line of the triplet is masked during the training and sampling processes since our goal is to

generate cues. Once trained, infilling is performed by using the preceding and succeeding dialogue lines as inputs to the model.

- **PPLM+LDA**: We extract keywords using LDA and control the generation process based on the topic of the dialogue.
- **PPLM+CueDisc**: We train a cue/dialogue sentence type discriminator and control the generation process using this classifier.
- **PPLM+Emotion**: We train a multi-label emotion classifier and steer the generation process to generate text that reflects the target emotion specified by the user.

**Parameter setup** For the PPLM experiments, we use the official PyTorch implementation published by the authors with modifications and extensions. We use the same parameters as PPLM to fine-tune the GPT-2 model. For all PPLM experiments, we set the step size  $\alpha$  to 0.04, the scaling coefficient for the normalization term  $\gamma$  to 1.0. Additionally, we keep the default values for the KL coefficient  $\lambda_{KL}$  and the gamma scale  $\gamma_{gm}$ , which are 0.01 and 0.95 respectively. The number of update steps  $m$  is 1 for all experiments, as we found that a larger number of update steps leads to more deterministic results. For ILM experiments, we train an ILM model with the default fine-tuning parameters specified in the Transformers library (see Wolf et al. [2019]), except that we use a batch size of 24 and a sequence length of 256. For all model experiments, we use a seed value of 0 and perform inference on a single GPU.

#### 4.5 Quantitative Results

We use n-gram similarity and distance metrics (see Kondrak [2005]) to measure the similarity of the generated text to our reference cue corpus, which consists of 50,000 cue samples from our training set. We generate 600 samples with each model and determine the top 10 reference cues for each sample that yield the smallest Levenshtein distance to the generated text. The Levenshtein distance is defined as the minimum number of elementary edit operations required to transform one string to another. We then compute the unigram and bigram similarity (LCSR and BI-SIM) for each generated sample and closest reference cue pairs, and report the average similarity over all generated samples. **PPLM+Emotion** and **PPLM+LDA** samples are generated using randomly selected target emotions and topics respectively. As shown in Table 2, PPLM with the Dialogue/Cue discriminator (denoted as **PPLM+CueDisc**) achieves the highest LCSR and BI-SIM scores, indicating that **PPLM+CueDisc** can successfully generate cues. The **PPLM+Emotion** and **PPLM+LDA** approaches achieve the second and third best LCSR and BI-SIM scores respectively. Since the **PPLM+Emotion** approach aims to generate text solely based on target emotion rather than sentence type (dialogue/cue), the results suggest that the dataset relies heavily on cues to convey target emotions, and that **PPLM+Emotion** therefore generates cues rather than dialogue lines. The fine-tuned GPT-2 medium model (denoted **GPT-2+ FT**) and the Infilling by Language Modeling (denoted **ILM**) have low LCSR and BI-SIM scores, suggesting that they are unable to generate relevant and complex cue structures.

Table 2: LCSR and BI-SIM scores of the models. PPLM+CueDisc shows the best performance in terms of LCSR and BI-SIM metrics.

Method	LCSR $\uparrow$	BI-SIM $\uparrow$
GPT-2+FT	0.42	0.29
ILM	0.47	0.24
PPLM+CueDisc	<b>0.72</b>	<b>0.60</b>
PPLM+LDA	0.68	0.55
PPLM+Emotion	0.69	0.57

In addition, we measure the diversity of the text generated by each model by the number of distinct n-grams (normalized by the length of text) as in Li et al. [2016]. We report the Dist-1, Dist-2, and Dist-3 scores for the distinct 1-2-3-grams in Table 3. As can be seen in Table 3, **PPLM+CueDisc** and **PPLM+Emotion** are comparable or better than **GPT-2+ FT** in generating diverse text while **ILM** and **PPLM+LDA** perform worst. On closer inspection, we find that some of the extracted cue keywords do not refer to characters, but to stage directions such as scene changes, lighting and sound instructions. Therefore, using the keywords extracted with LDA sometimes leads to the generation of repetitive, non-character related text. Similarly, we note that **ILM** also tends to generate repetitive

Table 3: Dist-1, Dist-2, Dist-3 scores of the models.

Method	Dist-1 $\uparrow$	Dist-2 $\uparrow$	Dist-3 $\uparrow$
GPT-2+FT	0.32	0.71	0.82
ILM	0.18	0.62	0.72
PPLM+CueDisc	0.25	0.69	0.80
PPLM+LDA	0.20	0.58	0.72
PPLM+Emotion	<b>0.34</b>	<b>0.74</b>	<b>0.87</b>

text that resembles stage directions. Some examples of scripts generated with (**GPT-2+ FT** and **PPLM+CueDisc**) can be found in Table 1. As can be seen from the examples, the GPT-2+ FT method is capable of generating plausible text, but not necessarily cues. In contrast, our method is able to generate cues with the characters that appear in the input text.

#### 4.6 Qualitative Results

We asked 20 human annotators to evaluate the performance of the models based on the coherence of the generated text and the accuracy of the cue generation. To create an evaluation dataset, we selected the best performing PPLM-based approach **PPLM+CueDisc** with the top competitor approach **GPT-2+ FT** and generated 50 random examples with each model. We then asked the evaluators to rate the generated examples based on coherence and cue accuracy in a binary manner. In the context of our work, we define coherence as both the independent plausibility of the generated text and the contextual coherence of the generated text with respect to the input sentence. Furthermore, we define cue accuracy as whether or not the text generated by the model contains a cue or not.

Table 4: Qualitative analysis with 20 human evaluators. The evaluators are asked whether the generated texts contain any cue (cue accuracy) and are coherent.

Method	Cue Acc $\uparrow$	Coherence $\uparrow$
GPT-2+FT	32.4	66.3
PPLM+CueDisc	<b>92.5</b>	<b>69.0</b>

As can be seen from Table 4, while **GPT-2+ FT** generates diverse text, it fails to generate cues given an input sentence. Since the majority of the dataset consists of dialogues, it is expected that the **GPT-2+ FT** approach is biased towards generates dialogues. On the other hand, our method achieves a high cue accuracy score while preserving the overall coherence of the conversation. However, we strongly believe that the coherence of the generated texts can be improved by better preprocessing steps and persona-based discriminators. We leave these ideas for future work.

## 5 Conclusion

In this paper, we use a large-scale play script dataset and propose the novel task of generating theatrical cues from dialogues. We approach the cue generation problem as a controlled text generation task and use a plug-and-play language model with a cue/dialogue discriminator, LDA-based topic keyword lists, and a multi-label emotion classifier to steer the language model to the desired attributes without re-training the model. Our experiments show that language models can be successfully used to generate plausible and attribute-controlled text in highly specialized domains such as plays. In the future, we plan to explore character and person-based cue and dialogue generation tasks with plug-and-play models.

## Acknowledgements

This publication has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TUBITAK (Project No:118c321). We also acknowledge the support of NVIDIA Corporation through the donation of the TITAN X GPU.



## References

- Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. *arXiv preprint arXiv:2010.01717*, 2020.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. Cue me in: Content-inducing approaches to interactive story generation. *arXiv preprint arXiv:2010.09935*, 2020.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- Elizabeth Clark, Yangfeng Ji, and Noah A Smith. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, 2018.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric C. Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *ArXiv*, abs/1912.02164, 2020.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Chris Donahue, Mina Lee, and Percy Liang. Enabling language models to fill in the blanks. *ArXiv*, abs/2005.05339, 2020.
- Stéphane d’Ascoli, Alice Coucke, Francesco Caltagirone, Alexandre Caulier, and Marc Lelarge. Conditioned text generation with transfer for closed-domain dialogue systems. In *International Conference on Statistical Language and Speech Processing*, pages 23–34. Springer, 2020.
- Bjarke Felbo, A. Mislove, Anders Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *ArXiv*, abs/1708.00524, 2017.
- HTGAA. How to generate (almost) anything project. <http://howtogeneratealmostanything.com>, 2017.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, R. Salakhutdinov, and E. Xing. Controllable text generation. *ArXiv*, abs/1703.00955, 2017a.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, R. Salakhutdinov, and E. Xing. Toward controlled generation of text. In *ICML*, 2017b.
- Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. Story generation from sequence of independent short descriptions. *ArXiv*, abs/1707.05501, 2017.
- N. Keskar, B. McCann, L. R. Varshney, Caiming Xiong, and R. Socher. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858, 2019.
- Grzegorz Kondrak. N-gram similarity and distance. In *SPIRE*, 2005.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL*, 2016.
- Fuli Luo, Damai Dai, Pengcheng Yang, Tianyu Liu, Baobao Chang, Z. Sui, and Xu Sun. Learning to control the fine-grained sentiment for story ending generation. In *ACL*, 2019.

- Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011.
- Sanidhya Mangal, Poorva Joshi, and Rahul Modak. Lstm vs. gru vs. bidirectional rnn for script generation. *ArXiv*, abs/1908.04332, 2019.
- Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a.
- Lara J. Martin, Prithviraj Ammanabrolu, W. Hancock, S. Singh, B. Harrison, and Mark O. Riedl. Event representations for automated story generation with deep neural nets. In *AAAI*, 2018b.
- Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, 2018.
- R. Plutchik. Emotion, a psychoevolutionary synthesis. 1980.
- Shrimai Prabhumoye, A. Black, and R. Salakhutdinov. Exploring controllable text generation techniques. *ArXiv*, abs/2005.01822, 2020.
- Alec Radford. Improving language understanding by generative pre-training. 2018.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- Mark Owen Riedl and Vadim Bulitko. Interactive narrative: An intelligent systems approach. *Ai Magazine*, 34(1):67–67, 2013.
- Rudolf Rosa, Ondřej Dušek, Tom Kocmi, David Mareček, Tomáš Musil, Patrícia Schmidtová, Dominik Jurko, Ondřej Bojar, Daniel Hrbek, David Košťák, et al. Theatre: Artificial intelligence to write a theatre play. *arXiv preprint arXiv:2006.14668*, 2020.
- Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Towards controllable biases in language generation. *ArXiv*, abs/2005.00268, 2020.
- Felix Stahlberg, James Cross, and Veselin Stoyanov. Simple fusion: Return of the language model. *ArXiv*, abs/1809.00125, 2018.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215, 2014a.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014b.
- Andrea Vanzo, Emanuele Bastianelli, and Oliver Lemon. Hierarchical multi-task natural language understanding for cross-domain conversational ai: Hermit nlu. *arXiv preprint arXiv:1910.00912*, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.
- Ke Wang and Xiaojun Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, 2018.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.
- Congying Xia, Chenwei Zhang, Hoang Nguyen, Jiawei Zhang, and Philip Yu. Cg-bert: Conditional text generation with bert for generalized few-shot intent detection. *arXiv preprint arXiv:2004.01881*, 2020.

- Yutao Zhu, R. Song, Zhicheng Dou, J. Nie, and Jin Zhou. Scriptwriter: Narrative-guided script generation. In *ACL*, 2020.
- D. Ziegler, Nisan Stiennon, Jeffrey Wu, T. Brown, A. Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593, 2019.
- A. Çelikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799, 2020.