DialogueScript: Using Dialogue Agents to Produce a Script

Anonymous ACL submission

Abstract

We present a novel approach to generating scripts by using agents with different personality types. To manage character interaction in the script, we employ simulated dramatic networks. Automatic and human evaluation on multiple criteria shows that our approach outperforms a vanilla-*GPT2*-based baseline. We further introduce a new metric to evaluate dialogue consistency based on natural language inference and demonstrate its validity.

1 Introduction

The last couple of years have seen some promising advancements in the area of open-ended story generation (Fan et al., 2018; Clark et al., 2018; Ammanabrolu et al., 2019), notably with the use of large pretrained generative neural language models such as GPT-2 (Radford et al., 2019; See et al., 2019). However, these works mostly focus on producing very short stories, such as those in ROCstories (Mostafazadeh et al., 2017). While there have been attempts at generating full-length theatrical works involving longer dialogue scripts, they use human-in-the-loop approaches, such as post-editing (Colton et al., 2016; Helper, 2018) or human choice between alternatives during the generation process (Rosa et al., 2021). Longer texts fully generated by language models (Sharp et al., 2016) often show as inconsistent and/or dull.

In this work, we explore a novel approach to generating longer scripted dialogues, such as theatre or movie scripts, inspired by works in personalizing dialogue agents (Zhang et al., 2018; Mazaré et al., 2018). Instead of handcrafting specific personas such as these previous works, we propose to cluster personalities based on major personality traits, i.e., the prevailing sentiment in the respective characters' utterances. We use these clusters to train three distinct models, which then act as a positive, neutral and a negative character. Since there are more than two characters, we need a non-trivial dialogue management system do decide the order of characters in the dialogue. We design a novel approach based on simulating dramatic networks (DN; Moretti, 2020). We compare our overall script generation approach to a baseline based on a vanilla GPT-2 model (Radford et al., 2019). We use basic automatic metrics for diversity and sentiment, combined with human evaluation on multiple criteria. Since automatic metrics for evaluating coherence of open-ended text generation are scarce, we present a new automatic metric based on natural language inference (NLI; Williams et al., 2018).

Our contributions include: (1) DialogueScript – script generation with distinct language models for different characters, based on character clustering; (2) dialogue management based on DN; (3) NLI-Score – a novel metric for the evaluation of consistency of the generation outputs; and (4) automatic and human evaluation comparing our DialogueScript/DN approach to a strong GPT-2 baseline. We plan to release our experimental code and models on GitHub.[1]

2 Script Generation Approach

2.1 Character Clustering

The characters in movies usually display a consistent personality within their utterances. However, training models for specific characters would make it difficult to explore various genres or situations due to training data sparsity. To find an acceptable balance between consistency and versatility, we simplify the training and group characters into several disjoint subsets based on their personality types. This selection is realized by a sentiment classifier by (Barbieri et al., 2020),[2] which is based on a pre-trained *RoBERTa-base* model (Liu et al., 2019), further trained on masked language modeling on on 58M tweets and finetuned on tweet senti-

[1] Link will be provided in the final version of the paper.

[2] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

ment classification. The model classifies the input into three groups, labeling it as positive, neutral or negative. Because the input length of the model is limited, processing all utterances of a character glued together would cause an undesirable input truncation. To address this issue, we label each dialogue turn individually and the overall character cluster assignment is computed as the prevailing sentiment over all their utterances.

2.2 Data Preprocessing for Language Models

Before the individual character language models are trained, the dataset needs to be pre-processed. Since the characters are identified by their membership in a cluster instead of their names, the name of each character is replaced by the label *focus* or *other*. The former denotes that the type of this character matches the sentiment of the model, e.g. when training a positive model, a positive character is labeled as the focus. The latter is used for marking characters that are not salient for current learning, e.g. the label of the negative and neutral characters in training data for the positive model. Because multiple characters within the same cluster may occur in one dialogue, several instances of every dialogue with different focus/other labels are included in the training data.

2.3 Simulating Dramatic Networks

To orchestrate the script generation between the separate character language models generating individual utterances, we design a new approach based on DN (Moretti, 2020). We consider the script/dialogue to consist of one or more exchanges (one character starting and others replying) and each line to be addressed to one specific character (i.e., character A utters a line addressed to character B). The dialogue flow is determined by interpretable parameters of characters and their relations. There are 3 main parameters per character:

- *centrality* – the probability of addressing another character (starting an exchange),
- *loyalty* – probability distribution over potential addressees,
- *reciprocity* – probability of replying to an address.

All parameters are updated throughout the script generation. Unlike Moretti (2020), we do not estimate model parameters from existing play scripts. Instead, we set initial model parameters empirically based on a few test trials, and we use the DN model to manage generation of new scripts.[3]

While all characters initially have the same centrality (i.e., the probability of starting the dialogue, set at 1), centrality increases with every line spoken by the given character. At the end of the script, each character's centrality reflects their significance for the generated script.

The loyalty parameter works similarly – if character A addresses character B at a given point in the script, their probability of addressing B in the future increases (at the expense of other characters). At the end of the script, the loyalty probability distribution reflects relationships a certain character had with all other characters. We set the loyalty probability distribution uniformly.

The reciprocity parameter determines if B responds to A after being addressed. To present a realistic length of exchanges between two characters in the script, reciprocity starts at 95% and decays by a third after each line uttered. The initial value and the decay rate are defined separately for each character. They determine the length of exchanges between two certain characters and reflect characters' talkativeness. Reciprocity resets after the end of a given exchange (when B decides not to respond to A). When an exchange ends, the next character to speak is chosen by centrality.

The probability of the dialogue ending after each line is independent of characters' relations; it is fixed at 20% throughout the generation.

3 Evaluation Metrics

Since standard reference-based language generation metrics are not applicable to our free-form long-text generation scenario, we combine basic corpus-based statistics showing diversity with evaluation of personality consistency via sentiment classification, coupled with human evaluation based on multiple criteria. We also propose a new automatic metric targeted at consistency.

3.1 Automatic Metrics

Diversity We evaluate several automatic metrics aimed at text diversity (van Miltenburg et al., 2018). This includes the perplexity, the total number of words generated, as well as the number of distinct words (1-grams) and bigrams. All diversity metrics are measured as average over generated dialogues.

[3]Moreover, while Moretti (2020)'s approach considers multiple scenes, we only assume a single scene/dialogue for simplicity.

Personality consistency To show that our models can generate consistent utterances based on the target character types, we measure sentiment of utterances in the generated dialogues, similarly to the training data clustering approach from Section 2.1.

3.2 Human Evaluation

We design two manual evaluation procedures, both to be carried out on the same text samples to reduce annotator mental load:

Relative ranking The annotators are asked to order dialogues generated by different systems from best to worst, according to their own subjective judgement, with no further instructions. This ranking gives us an overall system comparison.

Absolute scoring The annotators are asked to rate the generated dialogues in terms of the following properties on a 5-point Likert scale:

- *Coherence*: Is the text coherent?
- *Consistency*: Are the characters self-consistent?
- *Originality*: Is the text original and interesting?
- *Overall impression*: Did you enjoy reading this text?

3.3 NLI-Score: A Consistency Metric

Inspired by previous approaches using NLI to evaluate texts for other NLG tasks (Dziri et al., 2019; Maynez et al., 2020), we develop NLI-Score, a new metric for dialogue consistency.

In general, NLI determines whether a given sentence is entailed in, neutral to, or in contradiction with a context (Bowman et al., 2015). Unlike previous works, we aim at the *neutral* relation in NLI-Score, which indicates newly added information, but no inconsistencies. The contradiction relation indicates inconsistencies and the entailment relation is mostly indicative of repetition, both of which are unwanted in creative text generation. We use the *RoBERTa-large-mnli* model by (Liu et al., 2019)[4] to compute probabilities of the different NLI classes, then take the probability of the neutral category as the basis our NLI-Score. To make the metric robust to varied length, we propose to measure the average neutrality per added sentence. The second sentence is compared with the first, the third with the first two, and so on.[5]

[4] https://huggingface.co/roberta-large-mnli

[5] The context is truncated from the start if its length exceeds the NLI model's maximum input length.

4 Experiments

4.1 DialogueScript Training

In DialogueScript, characters are represented by three separate language models trained by fine-tuning the *GPT2-small* model (Radford et al., 2019), given the respective clustered data (positive, neutral, or negative) as described in Sections 2.1 and 2.2. The training uses an adaptive learning rate optimizer ($\alpha = 3 \times 10^{-5}$, $\epsilon = 1 \times 10^{-8}$) (Kingma and Ba, 2015) and a linear scheduler with warmup of 1,000 steps over five epochs.

To finetune the models, we use a dataset consisting of movie scripts (1,276 movies) from ScriptBase (Gorinski and Lapata, 2018) and TV show scripts (786 episodes) scraped from fan-sourced collections, IMSDb[6] and Forever Dreaming.[7]

4.2 Compared Model Variants

We evaluate 3 model variants: (1) a base *DialogueScript* model with random order of characters, (2) an extended *DialogueScript + DN* (based on the DN orchestration described Section 2.3), and (3) a *Baseline* based on vanilla *GPT2-medium* for comparison. Every generated dialogue includes three characters (each supposedly corresponding to one character type, i.e. positive, neutral and negative).

Both DialogueScript setups receive no textual initialization and generate scripts from scratch. This is not possible with the baseline, which requires a prompt to generate a script-like text.[8] Therefore, we use minimal prompts (a short 1-sentence setting description + single-utterance greeting from all three characters) to start the baseline model generation. These prompts are not included in the evaluation.

Note that the DialogueScript and DialogueScript + DN systems differ only in the order of the characters' utterances and the length of scenes. The dialogue management does not influence the content of the utterances themselves in any way, their content is generated using the same sentiment-based models (see Section 4.1).

4.3 Results

Automatic metrics For automatic evaluation, we use 50 scripts generated by our systems and 10 scripts by the *GPT2-medium* baseline. Table 1

[6] https://imsdb.com/

[7] https://transcripts.foreverdreaming.org/

[8] In our experiments, the unprompted *GPT-2* model generated HTML code.

Model	Perplexity	1-gram Vocab	2-gram Vocab	Words	NLI-Score
Baseline	1.86	59.25	88.00	104.00	0.40
DialogueScript	**2.48**	**241.90**	**428.06**	**489.76**	**0.47**
DialogueScript + DN	2.13	183.57	308.13	359.61	0.46

Table 1: Automatic metric results: generated script diversity (average perplexity, unigram and bigram vocabulary size, number of words) and consistency in terms of NLI-Score (see Section 3.3).

Character	Sentiment Positive	Neutral	Negative
Positive	35	64	1
Neutral	1	14	1
Negative	4	56	40

Table 2: Sentiment of the generated utterances, depending on the target sentiment for a given character.

Model	1st	2nd	3rd
Baseline	4	2	6
DialogueScript	2	5	5
DialogueScript + DN	6	5	1

Table 3: Results of relative ranking of model outputs.

Model	Coh	Con	Orig	Overall
Baseline	2.3	2.7	2.5	2.5
DialogueScript	**3.3**	3.2	3.8	3.3
DialogueScript + DN	3.0	**3.3**	**4.7**	**3.8**

Table 4: Average absolute human rating scores – Coherence, Consistency, Originality and Overall impression, on a 5-point Likert scale.

shows that both DialogueScript setups produce more diverse scripts than the baseline. Table 2 then demonstrates the inclination of DialogueScript model outputs to their target sentiment, with the exception of a prominent neutral sentiment. This is natural, because we cannot expect the characters to avoid common phrases with a neutral sentiment.

Human evaluation We use 12 short excerpts from scripts generated by each model for all of the manual evaluation tasks. The annotators are shown 5-10 lines[9] at a time. Each annotation is performed by 3 judges.

Table 3 with relative ranking results shows that DialogueScript + DN was most frequently the best option and least frequently the worst one. As we can see in Table 4, both our systems beat the baseline in all of the absolute scoring criteria. The DialogueScript + DN setup scores better than base DialogueScript with random character ordering on all criteria except Coherence. Since both DialogueScript setups use the same models, we believe that the DN orchestration made a difference in making the character interaction more organic.

NLI-Score We evaluated our new metric by comparing it to human evaluation of consistency. The scores have a Pearson correlation of 0.50, showing that NLI-Score does provide some consistency information. When we apply NLI-Score for automatic evaluation of the compared setups (see Table 1), we can see that NLI-Score is similar for both DialogueScript approaches and in both cases higher than the baseline, showing that our generated texts contain less detectable contradictions and repetitions than the baseline.

[9]The amount of text was similar for all evaluated dialogues as the number of lines was balanced by their length.

4.4 Discussion

While metrics such as perplexity can characterize an NLG output, they are not enough to decide on the overall output quality. However, we can use these characteristics to make an observation that our systems tend to be more verbose than the baseline approach. We hypothesize that this might have played a role in the human evaluation, especially in the ranking task where the baseline texts appeared sleeker and therefore easier to read.

5 Conclusion

We approached script generation by simulating the interaction of characters. We prepared training data for three different personality types (positive, neutral and negative) by clustering average sentiment values of characters in movies and TV shows. We trained the corresponding models and combined them by simulating dramatic networks. We proposed a new metric, the NLI-Score, to automatically evaluate the consistency of the generated text. Based on both automatic metrics and human evaluation, our approach outperforms the baseline in all of the observed qualities; our NLI-Score metric shows as indicative of overall output consistency.

References

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, et al. 2019. Guided Neural Language Generation for Automated Storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 46–55, Florence, Italy.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of EMNLP*, pages 1644–1650.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642, Lisbon, Portugal.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural Text Generation in Stories Using Entity Representations as Context. In *Proceedings of NAACL-HLT*, pages 2250–2260, New Orleans, Louisiana.

Simon Colton, Maria Teresa Llano, Rose Hepworth, et al. 2016. The Beyond the Fence musical and Computer Says Show documentary. In *Proceedings of the Seventh International Conference on Computational Creativity*.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating Coherence in Dialogue Systems using Entailment. In *Proceedings of NAACL-HLT*, pages 3806–3812, Minneapolis, Minnesota.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of ACL*, New Orleans, LA, USA.

Philip John Gorinski and Mirella Lapata. 2018. What's this movie about? A joint neural network architecture for movie content analysis. In *Proceedings of NAACL-HLT*, pages 1770–1781, New Orleans, Louisiana.

Roslyn Helper. 2018. Lifestyle of the Richard and family. https://www.roslynhelper.com/lifestyle-of-the-richard-and-family.

Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic gradient descent. In *Proceedings of ICLR*.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of ACL*, pages 1906–1919.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of EMNLP*, pages 2775–2779, Brussels, Belgium.

Franco Moretti. 2020. Simulating dramatic networks: Morphology, history, literary study. *Journal of World Literature*, 6(1):24 – 44.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Rudolf Rosa, Ondřej Dušek, Tom Kocmi, et al. 2021. THEaiTRE: Artificial intelligence to write a theatre play. In *Proceedings of AI4Narratives — Workshop on Artificial Intelligence for Narratives*, pages 9–13, Yokohama, Japan.

Abigail See, Aneesh Pappu, Rohun Saxena, et al. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of CoNLL*, pages 843–861, Hong Kong, China.

Oscar Sharp, Ross Goodwin, and Benjamin. 2016. Sunspring – A Sci-Fi Short Film Starring Thomas Middleditch.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the Diversity of Automatic Image Descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Saizheng Zhang, Emily Dinan, Jack Urbanek, et al. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL*, pages 2204–2213, Melbourne, Australia.