Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation

Judith van Stegeren
j.e.vanstegeren@utwente.nl
University of Twente
Enschede, The Netherlands

Jakub Myśliwiec
jmjmkuba@gmail.com
University of Twente
Enschede, The Netherlands

ABSTRACT

GPT-2, a neural language model trained on a large dataset of English web text, has been used in a variety of natural language generation tasks because of the language quality and coherence of its outputs. In order to investigate the usability of GPT-2 for text generation for video games, we fine-tuned GPT-2 on a corpus of video game quests and used this model to generate dialogue lines for quest-giver NPCs in a role-playing game. We show that the model learned the structure of quests and NPC dialogue, and investigate how the temperature parameter influences the language quality and creativity of the output artifacts. We evaluated our approach with a crowdsource experiment in which human judges were asked to rate hand-written and generated quest texts on language quality, coherence and creativity.

CCS CONCEPTS

• **Applied computing** → **Computer games**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

Natural language generation, quest generation, procedural content generation for games, MMORPG, World of Warcraft, GPT-2, Transformers, NPC dialogue, English

ACM Reference Format:
Judith van Stegeren and Jakub Myśliwiec. 2021. Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. In *The 16th International Conference on the Foundations of Digital Games (FDG) 2021 (FDG'21), August 3–6, 2021, Montreal, QC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472538.3472595

1 INTRODUCTION

The Transformer architecture [26] is a recent breakthrough in the NLP research field. Transformer-based language models, like GPT-2 [21] and BERT [9], represent the state-of-the-art in various language processing and generation tasks. We want to apply the Transformer architecture to procedural content generation for video games, to see how it can benefit researchers, game developers, publishers and players.

In this paper we investigate the efficacy of Transformers for dialogue generation for quests in role-playing video games. Role-playing games are a particularly suitable application domain for natural language generation techniques, as this genre leans heavily on narrative [2]. In many RPGs, the main game narrative is supplemented with side-quests, i.e. non-obligatory small errands that exist to give the player more freedom [20]. Side-quests tend to follow a fixed structure, so they lend themselves well to procedural generation. We use the massive multiplayer online RPG World of Warcraft [3] as a case study for our research.

In World of Warcraft, quests have a title and a quest objective, i.e. the assignment that the player should complete to earn the quest reward, such as experience points, items or money. Typically, quests are obtained by the player through special quest-giver NPCs. These NPCs introduce the quest to the player and contextualize it in the game world with a few lines of dialogue. The dialogue lines with the quest's backstory are a form of flavor text, i.e. decorative text that is not essential to the gameplay. If we remove the dialogue lines from the game, the quest is still playable, but the quest is no longer explicitly linked to the game world and the game narrative. We have investigated whether GPT-2 can be taught the lore of World of Warcraft, and whether we can use this model to generate flavor text for quests, given a title and an objective. There are already many systems which can generate video game narratives and quests in certain structures [15]. However, most of these systems use rule-based systems to generate surface text for those stories [16]. We explore whether GPT-2 is a viable alternative for rule-based approaches, especially for generating quest surface text.

Our contribution is two-fold. We have created a fine-tuned GPT-2 model based on a annotated dataset of World of Warcraft quests. We then show that this model can be used to generate new dialogue lines for quest givers in World of Warcraft, given a human-written quest objective and a quest title as a prompt. We have evaluated our approach by comparing hand-written texts from the same game with the outputs of our generator on the properties of language quality, coherence and creativity in a crowdsourcing experiment. Our code and dataset are available for re-use by other researchers, and a working demo of the trained model can be found online.

2 RELATED WORK

Our research is related to work that investigates neural language generation with GPT-2, and procedural content generation for video games. In the latter category, we build on quest generation and dialogue generation research.

FDG'21, August 3–6, 2021, Montreal, QC, Canada

ACM ISBN 978-1-4503-8422-3/21/08.
https://doi.org/10.1145/3472538.3472595

2.1 Neural language generation

In 2018, OpenAI released GPT-2 [21], a language model based on the Transformer [26] architecture. The Transformer is currently one of the best performing general language models for natural language generation tasks, such as machine translation. GPT-2, like its commercial successor GPT-3 [5], is one of the largest models of this type. With additional fine-tuning on smaller, more specific datasets, GPT-2 can be trained to perform a variety of natural language processing tasks. GPT-2 has been pre-trained on a large corpus of web text, which results in a language model with robust latent knowledge of the English language. It is capable of generating arbitrary text which is largely grammatically correct, and its output is coherent enough that some people have difficulty identifying whether the text has been written by a computer or a human.

Global coherence, as opposed to coherence on the sentence and paragraph level, is still a problem for many neural text generation systems. [19] For text generation for games, this problem is even more pronounced, as output text should not only be coherent at the global text level, but also be coherent with the game. Various authors have investigated ways to improve the coherence of generated texts, for example by using planning strategies [13, 29], learning frameworks [12] and coherence metrics [19]. Some generation systems do not actually enforce coherence in their outputs, but only suggest it through their form or content to trick the reader [11, 24].

When Transformer-based architectures are used for generation, various authors use special tags, tokens or markers [14, 17, 30] to annotate their training data with additional information. The special tags can be used to describe the structure of the training data, or draw attention to specific parts of the input. In other words, the special tags change how the language model reads the training data. When the language model has been fine-tuned on annotated training data, it can be guided towards a certain output by reusing the special delimiter tokens in the generation prompt. Guiding the generator, for example towards using certain words or themes, can be used to increase the perceived coherence of the output. This can be coherence at the global text level, or coherence between the output and its larger context, such as a video game narrative.

A notable example of neural text generation for video games is the game AI Dungeon [28]. AI Dungeon acts as a cross-over between a virtual tabletop Dungeon Master and a classic text adventure game. Text adventure games generally consist of a pre-written narrative, and accept input from the player in the form of natural language text, such as 'open door', 'pick up sword' or 'fight grue'. AI Dungeon is similar to these classic text adventures in that it accepts natural language input from the player. However, instead of following a pre-written story, a neural language model extends the story on the fly, using the user input as prompt. Earlier versions of AI Dungeon used GPT-2[1] fine-tuned on text from Choose your own adventure-style games.

AI Dungeon's underlying neural model has some similarities to our approach: both are a fine-tuned version of the pre-trained GPT-2 model, trained on structured data from narrative games. However, AI Dungeon's approach differs from ours in that it was finetuned on a more generic dataset, spanning different language styles, narrative genres and topics. Our GPT-2 model was trained on homogeneous data, namely quests from World of Warcraft's game setting. Furthermore, the goals of both language models are fundamentally different. AI Dungeon is meant to function similarly as a Dungeon Master. In table-top games, Dungeon Masters should be able to respond to any 'input' that the players throw at them, within the rules of the game. As a result, AI Dungeon's language model should be able to deal with a huge variety of inputs. By design, the model can be used for text generation without any human supervision; text is generated, and immediately presented to the player. Our language model, on the other hand, is meant as an authoring aid for game writers. The idea is that generative language models can be used to do the heavy lifting in creating new content, something that may benefit game development companies as the playable content for video games, especially for open-world RPGs and MMORPGs, increases. Outputs are not meant for direct inclusion in a game; instead, they should be checked and possibly modified by a human writer.

[1] https://pcc.cs.byu.edu/2019/11/21/ai-dungeon-2-creating-infinitely-generated-text-adventures-with-deep-learning-language-models/

2.2 Quest generation

Kybartas and Bidarra [15] discuss various examples of quest generation research in their survey paper about story generation. One of the early approaches is SQUEGE [20]. It relies on patterns it randomly selects and then populates the blanks with appropriate information about characters, locations and items. The authors of ReGEN [16] built on SQUEGE's approach and proposed a method for narrative graph rewriting, which is capable of creating complex branching stories. However, the surface text of the generated quests is still created with a rule-based approach, which limits the variety of the language of the final artifact. Doran and Parberry [10] explored the use of context-free grammars in RPG quests. As rule-based models are generally not generalizable to other domains, these approaches risk becoming repetitive.

Ammanabrolu et al. [1] compared two generation approaches for quest generation in a text-based cooking game: one based on Markov chains, and one based on neural generation. The authors used a fine-tuned GPT-2 model to generate surface text, i.e. cooking instructions, for their generated quests. Crowdsource workers were asked to play two generated quests and evaluate the quests on the properties of perceived creativity and coherence. The authors found that neural generation offers more value and greater coherence, whereas the Markov model produced quests that are more surprising and novel. They also noted that the neural generation approach required less domain knowledge and is potentially more generalizable to other domains.

2.3 NPC dialogue generation

In World of Warcraft, the flavor text for quests consist of quest-giver NPC dialogue. Other researchers have investigated techniques for dialogue generation for games, although we have not yet seen dialogue generation for games based on Transformers such as GPT-2. Walker et al. [27] generated dialogue for the prototype role-playing game SpyFeet, based on statistical machine learning models trained on film character dialogue. Ryan et al. [23] presented a method for annotating human-authored dialogues, so that different parts of these dialogues can be recombined to form new dialogues. The

authors tested their approach on dialogue of the social AI engine system Comme il faut (CiF), used in the game Prom Week. Tracery [8], a tool for generative grammars, has been used in various games to generate in-game text, including dialogue. Ryan et al. [22] created Expressionist, an authoring tool for generating in-game text at runtime. The tool is based on context-free grammars with added markup labeling. Users can use user-chosen tags to annotate the non-terminal symbols with arbitrary metadata, which gives new expressive power to the context-free grammars. Lessard et al. [18] used Expressionist to generate dialogue for their resource management game Hammurabi.

3 METHOD

We want to build a text generator that, given a quest title and objective, can generate dialogue lines for the quest-giver NPC. Our approach is fine-tuning a pre-trained GPT-2 language model on a dataset with quest information, and using the fine-tuned model to generate NPC dialogue. In this section, we start by discussing our training data, which is followed by a description of how we fine-tuned the model and used it to generate new dialogue.

3.1 Data

When we started this research project, we had not chosen a particular game to work with yet. To fine-tune GPT-2's pre-trained model, we needed a dataset with quest data. Specifically, we wanted to finetune the model with training data that has been annotated with tags that describe the structure of the data, following the approach of Zellers et al. [30]. Consequently, we needed a dataset that included NPC dialogue, and quest titles or quest objectives to contextualize the NPC dialogue. We estimated that using data from Massive Multiple Online RPGS (MMORPGs) would be a good idea, since these generally have more playable content than offline games and single-player games. These games tend to have large player communities, which could mean that information about quests would be obtainable online, for example from fan-websites [25]. Initially, we identified three possible sources for a dataset that fulfills this requirement, with data from popular MMORPG games:

(1) WOWHead quest database[2] which contains 24,981 quests from World of Warcraft [3];
(2) Destiny tracker quest database[3], containing 680 quests from Destiny [6];
(3) EVEinfo.com mission database[4] with 310 EVE Online [7] missions

It is possible to use GPT-2 without any fine-tuning, or fine-tune GPT-2 with relatively small datasets. However, since we want to teach GPT-2 a particular structure (i.e. a quest consists of a title, objective and dialogue), the dataset should be large enough for the model to capture the structure of the training data. A dataset with a few hundred datapoints is probably not large enough to change the pre-trained model in a significant way, and thus to generate outputs that follow the structure of a quest. We estimated that, individually, the Destiny and EVE Online datasets were probably too small to use as training data. A possible solution for this is combining the three datasets into one large dataset. However, although their structure is the same, the three datasets are too different to combine successfully. EVE Online and Destiny are science fiction games, whereas World of Warcraft is a fantasy game. Destiny quests have very short quest descriptions in the form of a quote, as opposed to World of Warcraft and EVE Online quests, which have longer dialogues as quest description. EVE Online's quest objectives lack in variety, as quests fall in one of only four categories.

During preliminary testing, we found that fine-tuning on one homogeneous dataset leads to higher quality output. Since the Destiny and EVE Online datasets are relatively small (only a few hundred quests compared to WoW's 24,000 quests), we decided to use only the World of Warcraft quest database. Figure 1 shows an example quest from this dataset.

Title
The Wayward Crone
Objective
Confront Helena Gentle in her home outside of Fallhaven
Description
The ledger indicates that an old woman named Helena Gentle recently took up residence in a house down the road from the town. The villagers' writings point to her being involved somehow with a variety of maladies that struck the village recently. It's possible that she may know what's behind this spell, if it hasn't afflicted her as well.

Figure 1: Quest 'The Wayward Crone' from the World of Warcraft quest database

Structure

`<|startoftext|> [quest title]`
`<|obj|> [quest objective]`
`<|text|> [quest description]`
`<|endoftext|>`

Example datapoint
`<|startoftext|>` In Dire Need `<|obj|>` Hear out the Council of Six in the Purple Parlor. `<|text|>` The ill tidings you bear only increase the concerns the Council has been having. I know you have your hands full with the Tomb of Sargeras. Make no mistake, the battle for the tomb remains our top priority–but we must not think the Legion foolish enough to rely on brute force alone. They are far more cunning than that. Please, hear us out. `<|endoftext|>`

Figure 2: Structure of datapoints in our training set. Quests are annotated with special tags that denote the quest title, objective and description (NPC dialogue).

[2]https://www.wowhead.com/quests
[3]https://db.destinytracker.com/d1/quests
[4]https://eveinfo.com/missions/

3.2 Training

We added tags to the dataset that describe the structure of the datapoints. Figure 2 shows the tags we used to annotate the quests in the World of Warcraft database, together with a concrete example datapoint from our training set. GPT-2 uses these tags to learn the structure of the datapoints in the training set. During generation, we can use these tags to steer GPT-2 towards a particular output structure: if we provide a partial datapoint (title and objective) as a prompt, GPT-2 will expand it with the part we want to generate (quest description, i.e. dialogue). We used the second-largest GPT-2 language model with 774 million parameters, which we deployed in a Google Colab environment, a Jupyter notebook environment in the cloud that comes with computation time on GPU. The basis of our code is Max Woolf's gpt-2-simple[5] project. Fine-tuning GPT-2 on our annotated training data took approximately 4 hours using an NVIDIA Tesla V100 GPU.

3.3 Generation

We use a quest title and objective, together with their start tags, and the starting tag for NPC dialogue, i.e. `<|text|>`, as prompt for the generator. For an example prompt, see Figure 4. The generator creates new text by sampling the fine-tuned language model for follow-up tokens. The model continues generating text until it has generated a pre-determined maximum amount of tokens. The output is cut off at the first `<|endoftext|>` token.

We can influence generation by changing the temperature parameter. A higher temperature leads to more unexpected output, which influences output properties like coherence, language variety, interestingness and coherence. Figure 3 shows example outputs for various temperatures. Since temperature determines the predictability of outputs, choosing a too low temperature leads to repetitive language in the output text. The last example in Figure 3, generated with a temperature of 0.5, demonstrates this, as it consists mostly of words and phrases that already occurred in the prompt title. A temperature of 0.7 leads to better quality outputs than a temperature of 0.5. The generator created a quest description that is an exact copy of the quest objective, but some of the outputs also contain new phrases. The outputs with an even higher temperature of 0.9 are more unexpected and contain more variety, thus we choose this temperature for generating our evaluation outputs.

Qualitative inspection of the generated quests shows that the fine-tuned GPT-2 model outputs artifacts that are highly coherent with the prompt. In Figure 3, multiple generated quest descriptions contain references to 'the Council of Six', a name that appears in the quest objective of the prompt.

Interestingly, our model has learned the structure so well that it often generated entire quests by itself. They follow the exact same structure as our training data, i.e. title, objective, and description, with each part delimited by tokens. Because the training set was taken from World of Warcraft, the generated quests contain words, phrases and names that are reminiscent of the lore of the game world.

[5] M. Woolf. GitHub - minimaxir/gpt-2-simple: https://github.com/minimaxir/gpt-2-simple

4 EVALUATION

4.1 Experiment Design

We conducted an online survey to evaluate the outputs of the generator. Participants were presented with 20 quests, which were ordered randomly. 10 quests were randomly chosen from the World of Warcraft dataset, and 10 quests were generated with our fine-tuned GPT-2 model, using a randomly chosen title and objective from the training set as prompt. To keep the questionnaire brief, we decided to only evaluate outputs that were generated with a temperature of 0.9. The generated quests were generated in one go, and not cherry-picked from a larger set of outputs. Figure 5 lists the quests that were used in the evaluation survey. For each quest the participants were asked to indicate how much they agree with 5 statements on a 7-point Likert scale. Participants rated all 20 quests on 5 statements, for the properties language quality, coherence and creativity:

Language quality Participants were asked to indicate their agreement with the statement "The description makes use of correct English."

Coherence To measure the perceived coherence between the quest objective and the NPC dialogue, we used the statement "The goal is clear from the description."

Creativity We used Boden's breakdown of creativity, similar to the approach of Ammanabrolu et al. [1]. According to Boden [4], creativity can be measured in terms of novelty, surprise and value. We operationalized this property with three statements:

(1) "The description is written in a novel way."
(2) "The description is surprising."
(3) "The description is creative."

Based on preliminary testing of the survey questions we decided to substitute the word 'value' with the word 'creativity', as participants found the question about value too confusing.

4.2 Results

32 participants filled in the online survey. For each participant the scores were grouped into 2 categories: generated and hand-written. We calculated the average for each property, and conducted a sign test on the median as none of the differences showed a normal distribution. Figure 6 shows the average score per category for each of the five statements. Three properties, namely language quality, coherence and novelty, were shown to be statistically significantly worse in descriptions generated by the model, with $p < 0.01$. The two other properties, surprise and creativity, did not show significant differences even with $p < 0.05$.

Even though the generator on average performs worse than human writers, the results are encouraging. The scores of generated quest descriptions for surprise and creativity are close to those scores for the hand-written descriptions. This may be because of the high temperature with which the quest description were generated. Similarly, the differences in language quality and coherence might be due to the high unexpectedness of the generated texts, due to the high temperature setting used during generation.

	Temperature = 0.9
1	"The Council has a rallying point all across The Dreamgrove. There is an empty throne oak in the purple parlor, and a creature named Amalia should be there. Unfortunately her days of being known as a Gallywix-level concession vendor are done."
2	"Hear me out! We're in dire need of help. Our explosive ordnance is getting dangerously low on both of our runs. I know sylvari aren't usually known for their industriousness, but this barrel needs tolled and our prophet needs his prayers answered. Who knows, maybe you can help us out."
3	"I will not keep you waiting any longer. I can offer you an undersea vantage that will let you see Bevel's will done. Please, take advantage of this juxtaposition of nature and technology. Meet with the Council of Six in the Purple Parlor. When you are prepared, we will speak again."
4	"Greetings, and welcome to the Emerald Sanctum. It is with great pleasure that the Council of Six can present to you its members. Peerless Master of the Arena Shantelos, Peerless Trainer of Arena Mastery, Champion of the Lonely theurge, and Peerless Master of Evodyne Industries are bound by the will of the Council of Six. Do welcome your retinue to the circle of gladiators."
5	"The felbat is a vicious predator, and the last thing we need in the Broken Isles is another of our kind. Report to the highest ranks of the Kirin Tor and tell them of our plight. The Broken Isles is a dangerous place where only the fittest survive. I'll designate a short period of time for them to consider our concerns. I should take my leave. Stay well."
	Temperature = 0.7
6	"Your aid is needed here, not in the Garden, but in the Purple Parlor, where the Council of Six holds its meetings. Among those attending are... well, let's see. It wouldn't do for me to name them, would it? The ones that might know are... well... let's see... The Council of Six is worried about the new naaru. I can't leave them in this place. I need to go back to the Hall of the Guardian. Hear out the Council of Six in the Purple Parlor."
7	"Master Ashwall wants you to head to the Purple Parlor in Zoram'gar Outpost and inform them of what you have learned."
8	"Hear out the Council of Six in the Purple Parlor!"
9	"The Alliance needs your help! Many of our brothers and sisters have been placed in peril. Please hear out the Council of Six, we need to know how to handle this situation."
	Temperature = 0.5
10	"The Council of Six is a council of the six. The leader of the Council of Six is called the "Crown Prince." The Crown Prince has a council with a council. The Council of the Six is a council of the six. The Council of the Six is a council of the six. The Crown Prince is a king. The Council of the Six is a council of the six. The Council of the Six is a council of the six."

Figure 3: Quest descriptions generated by our finetuned GPT-2 model with temperatures ranging from 0.5 to 0.9. All outputs were generated with the prompt from Figure 4. We can see that a temperature of 0.5 is too low to create usable text.

```
<|startoftext|>In Dire Need<|obj|>Hear out the Council
of Six in the Purple Parlor.<|text|>
```

Figure 4: Prompt for the quest 'In Dire Need' from World of Warcraft. The prompt is used to generate NPC dialogue for the quest-giver NPC.

Figure 7 shows the distribution of the ratings across the different properties. For all properties, the generator is capable of generating outputs that score higher than the median rating for hand-written texts. Sometimes the quality of generated quest descriptions even exceeds that of handwritten examples, such as in the case of the properties coherence and surprise. The inconsistency in quality of GPT-2's outputs suggests that we should take a cherry-picking approach to generation. Instead of generating one description, we could use the system to generate multiple outputs for the same prompt, and select the best option. If we can come up with a automatic metric for judging the quality of generated artifacts, we could also use a automated generate-and-test approach. If the generator should be used without any input from the user, generating texts with a lower temperature (e.g. 0.7) might improve the language quality and coherence of the output, possibly at the detriment of creativity and surprise. As we can see in Figure 3, texts generated with a lower temperature seem to be more consistent in both spelling and informative content.

5 DISCUSSION AND CONCLUSION

Although the fine-tuned model creates outputs that score worse on all three properties, the described approach is a viable option for text generation for games. Teaching GPT-2 a specific output structure by delimiting the training set and prompts with unique tokens proved to be a success. Although the quality of the results may be slightly inconsistent across multiple outputs generated with the same prompt, the outputs indicate that GPT-2 can learn the structure and linguistic style of World of Warcraft quests. Moreover, the ratings distributions show that the generator is capable of

Title and objective	Quest description
Trouble at the Docks. Crane Operator Bigglefuzz in Ratchet wants you to retrieve the bottle of 99-Year-Old Port from Mad Magglish who is hiding in the Wailing Caverns.	My wife took me to the Docks last week to appear on one of those 'news' shows some goblin or goblin-looking creature is running. We were going to get a discount as it was a special occasion, but the first news report I saw, Mad Magglish was already in the caves and stealing all the time. He's the one who gave that old pirate outfit to his son, who is basically a glorified beer sailor now. It's never safe to take old pirates' clothes, they usually have valuable chestnuts inside.
Proof of Strength: Ogron Horn. Take the proof to Limbflayer at Beastwatch in Gorgrond.	Take this to Limbflayer. I can't guarantee that he'll believe it, but I'm confident that he can convince it's none of my business. I've seen the certain beasts growl, so perhaps he'll take it as a personal insult that I'm actually here. But I bet he won't. You've proven your mettle. You deserve the horn.
Rituals of Power. Bring the book Rituals of Power to Tabetha in Dustwallow Marsh.	Something in this book seems to be purging the memory of the demi-god, Rukhmar. I'm not sure what purpose it could be serving, but it's disabled and cannot be used again until the ritual is prepared. You'll need some materials if you plan on performing the ritual. Bring me a copy of the Rituals of Power and I'll create the ritual for you. It should only take you a few minutes. Travel to Orgrimmar and speak with my assistant, Tabetha.
Spray it Again. Bring the Box of Bugkill Fumigators to Marcy Curtainfire inside the Ban'Thallow Barrow Den.	There you have it. Of course the crystal started glowing before the attack. I will have to perform some further research in the laboratory, but first I'm going to need to go set the rest of the explosives. You seem to know where that stuff goes. It's unfortunately hanging out in a safe in the back of this barrow den. If you want me to retrieve it, all you have to do is give me your key. Of course, I won't charge you for the hassle.
The Bandit Warlord. Slay Warlord Ihsenn.	There is an enemy that exceeds even Neferset that the guards are too weak to handle, he is called Warlord Ihsenn. If you kill him, perhaps I can retake Felfire Hill.
A Well Rounded Defense. The Facet of Keanna stored inside Keanna's Will wants you to gather 20 Arcane Residues from the Arcane Anomalies in Karazhan.	Our people have been fighting an endless war against necromantic attacks, and we cannot afford to stagger their resistance. I want you to find and gather arcane residue from arcane anomalies throughout the city. I want you to take this branding iron and release the broken into the city to get a better look at the matter. You will soon find that they are everywhere, and you will find that they generate a powerful explosion.
The Road to Purgation. Accompany Kingslayer Orkus to Purgation Isle.	It's that time again. You've earned it. You can finish off these hereers before they make any strikes. Let's go party.
Balance of Power. Speak to Romuul's assistants to collect an Optical Disc, Conducting Crystal, and Romuul's Wrench.	Sound difficult? We ought to be able to get these books back to Kumul without too much trouble. Don't worry though, I'll make it worth your while. First, we need to hide one of the books. We can do that by giving it a nice, "Hey, you know what? I got this one figured out, go figure it out for me!" put down. Then we need to get the books off our chests before they get confiscated. There will be a bunch of assistants nearby to familiarize them with your face.
Fashion History and a Philosophy of Style. Learn from Eliezer Hammerbeard how to create "flair and drama" in your wardrobe.	My job as a tailor is to take your stuff and make it look good. The way I see it, there are two ways to get people to like you: 1. Show them how good you are and 2. Show them how good you are through stylishly made clothes. I will cover both in this class. Please come back to me after this class and I will demonstrate how to take ye best pieces and make ye clothes that people want to buy.
In Dire Need. Hear out the Council of Six in the Purple Parlor	The Council has a rallying point all across The Dreamgrove. There is an empty throne oak in the purple parlor, and a creature named Amalia should be there. Unfortunately her days of being known as a Gallywix-level concession vendor are done.

Figure 5: Generated quest descriptions that were used in the evaluation survey. The quest title and objective were selected randomly from the WoW dataset and used as a prompt for the generator. All outputs were generated with a temperature of 0.9**.**

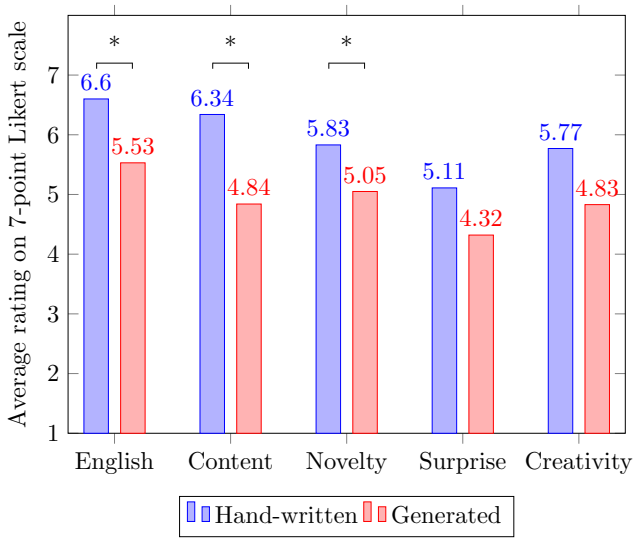

Figure 6: Average rating across evaluation properties rated on a 7-point Likert scale, for hand-written and generated quest descriptions. We collected ratings from 32 participants. The $*$ denotes statistical significance with $p < 0.01$.

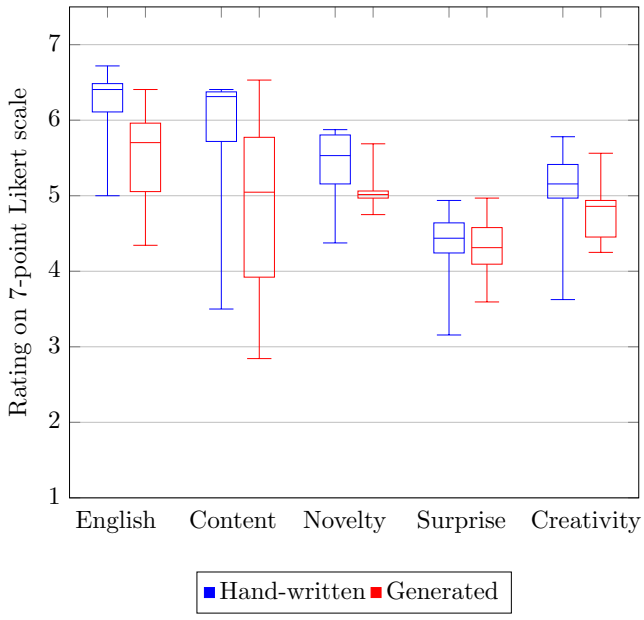

Figure 7: Distribution of ratings per evaluation property for hand-written and generated quest descriptions. 32 participants rated each property on a 7-point Likert scale.

creating outputs that are on par with human-written dialogues. A difference between human game-writers and our fine-tuned model is that the generator can easily create large numbers of quest descriptions from the same prompt. Once a GPT-2 language model has been fine-tuned, creating large volumes of embryonic quest descriptions is fast and low-cost. Letting a human user cherry-pick outputs with the highest quality, or modify the most creative outputs, seems a feasible alternative to writing new RPG quests by hand. Finding the optimal generation temperature could lead to additional improvements in output quality.

The weak point of this approach is that the fine-tuned GPT-2 model does not yet generalize to other role-playing games. Since we fine-tuned GPT-2 on World of Warcraft data, the resulting outputs contain references to the lore of World of Warcraft, and the outputs are not easily transferable to other game worlds.

However, we imagine the model can be adapted to generalize in a few ways. Firstly, we could try to create a more general model by preprocessing the training data: if we substitute all named entities by a `[LOCATION]`, `[NAME]` or `[FACTION]` tag prior to fine-tuning, the outputs will become a lot more generic. The placeholders can then be filled with names and location of another game world, either by hand or automatically. For the automatic substitution, we could use a language model that performs well on the cloze test.

We might also try to translate outputs from one game to another using techniques from neural machine translation in a post-processing step. For example, we could try to translate game specific terms using distributional semantics, i.e. by building a vector space of word embeddings using words from two different games. For example, if it turns out that 'Sith' (Star Wars games) and 'Undead' (World of Warcraft) are often used in the same context, these words might appear close together in the resulting vector space. We could use this information to substitute the word 'Undead' for the word 'Sith' in generated outputs.

We could also fine-tune GPT-2 on a heterogeneous dataset from different video games, annotated with additional tags that denote the game world or game genre that a datapoint originates from. These tags could then be used to steer the model towards outputs for a specific genre or game. A prerequisite for this approach is a varied dataset that is large enough to capture significant differences in style and content between games. Although nothing of this type was available when we performed this research, some efforts (e.g. [25]) are being made to compile such a dataset for subsequent text generation research.

Other future work in NLP for video games could explore generation in the opposite way than we did in this paper, i.e. take a piece of dialogue as prompt to generate a quest title and objective. This can be seen as a form of text summarization, as the quest objective should be grounded in the dialogue.

Evaluating the impact of different temperature setting in a more rigorous evaluation could also be useful, although it is doubtful whether these findings would easily generalize to models trained on other (game) data.

Following the approach of fine-tuning GPT-2 on a training set annotated with tags that describe the structure of input texts, we should explore whether we can add additional annotation tags to the training set, for example for capturing expressed sentiment and NPC-player relations. These tags could then be used in prompts, so that we can exercise more control over generation by guiding the generator towards outputs with desired properties. Using the largest GPT-2 model (1.5 billion parameters) might improve the language quality of the generated examples. However, if we start using larger pre-trained language models, we must also investigate whether the size of the training set should be increased proportionately, to

prevent the larger model from undertraining on the annotated game data. It would also be interesting to find out how large a dataset of game texts should be, before it can be used to teach GPT-2 the structure and linguistic style of game texts. One of our reviewers observed that some World of Warcraft quests contain references to pop-culture.[6] Automatically inserting memes, jokes and pop-culture references in generated dialogue is likely to happen, since GPT-2 was pre-trained on web text. If we fine-tune GPT-2 on a dataset of video game text from multiple games, the amount of references to other games will increase. As an added bonus, players that encounter the generated dialogue in-game might conclude that the references were introduced on purpose by human writers.

The system described in this paper is open source. The training data and generator used for the experiments can be found as a Google Colab environment at https://jakub.thebias.nl/research/QuestGen/colab/.

ACKNOWLEDGMENTS

This research is supported by the Netherlands Organisation for Scientific Research (NWO) via the Data2Game project (project number 055.16.114).

REFERENCES

[1] Prithviraj Ammanabrolu, William Broniec, Alex Mueller, Jeremy Paul, and Mark Riedl. 2019. Toward Automated Quest Generation in Text-Adventure Games. In *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 1–12. https://www.aclweb.org/anthology/2019.ccnlg-1.1

[2] Chris Bateman. 2007. *Game writing: Narrative skills for videogames*. Charles River Media.

[3] Blizzard Entertainment. 2004. *World of Warcraft*. Game [PC]. Blizzard Entertainment, Irvine, California, US.

[4] Margaret A. Boden. 2007. Creativity in a nutshell. *Think* 5, 15 (2007), 83–96. https://doi.org/10.1017/s147717560000230x

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Bungie. 2014. *Destiny*. Game [PlayStation, Xbox]. Bungie, Bellevue, Washington, US.

[7] CCP Games. 2003. *EVE Online*. Game [PC]. CCP Games, Reykjavík, Iceland.

[8] Kate Compton, Quinn Kybartas, and Michael Mateas. 2015. Tracery: an author-focused generative text tool. In *International Conference on Interactive Digital Storytelling*. Springer, 154–161.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[10] Jonathon Doran and Ian Parberry. 2011. A Prototype Quest Generator Based on a Structural Analysis of Quests from Four MMORPGs. In *Proceedings of the 2nd International Workshop on Procedural Content Generation in Games* (Bordeaux, France) *(PCGames '11)*. Association for Computing Machinery, New York, NY, USA, Article 1, 8 pages. https://doi.org/10.1145/2000919.2000920

[11] Jason Grinblat and C. Brian Bucklew. 2017. Subverting historical cause & effect: generation of mythic biographies in Caves of Qud. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (Hyannis, Massachusetts). ACM, ACM, New York, NY, USA, Article 76, 7 pages. https://doi.org/10.1145/3102071.3110574

[12] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to Write with Cooperative Discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1638–1649. https://www.aclweb.org/anthology/P18-1152

[13] Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 329–339.

[14] Tassilo Klein and Moin Nabi. 2019. Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. *arXiv preprint arXiv:1911.02365* (2019).

[15] Quinn Kybartas and Rafael Bidarra. 2017. A Survey on Story Generation Techniques for Authoring Computational Narratives. *IEEE Transactions on Computational Intelligence and AI in Games* 9, 3 (9 2017), 239–253. https://doi.org/10.1109/TCIAIG.2016.2546063

[16] Quinn Kybartas and Clark Verbrugge. 2014. Analysis of ReGEN as a graph-rewriting system for quest generation. *IEEE Transactions on Computational Intelligence and AI in Games* 6, 2 (2014), 228–242.

[17] Jieh-Sheng Lee and Jieh Hsiang. 2020. PatentTransformer-2: Controlling Patent Text Generation by Structural Metadata. *arXiv preprint arXiv:2001.03708* (2020).

[18] Jonathan Lessard, Etienne Brunelle-Leclerc, Timothy Gottschalk, Marc-Antoine Jetté-Léger, Odile Prouveur, and Christopher Tan. 2017. Striving for Author-Friendly Procedural Dialogue Generation. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (Hyannis, Massachusetts) *(FDG '17)*. Association for Computing Machinery, New York, NY, USA, Article 67, 6 pages. https://doi.org/10.1145/3102071.3116219

[19] O.O. Marchenko, O.S. Radyvonenko, T.S. Ignatova, P.V. Titarchuk, and D.V. Zhelezniakov. 2020. Improving Text Generation Through Introducing Coherence Metrics. *Cybernetics and Systems Analysis* 56, 1 (2020), 13–21.

[20] Curtis Onuczko, Duane Szafron, Jonathan Schaeffer, Maria Cutumisu, Jeff Siegel, Kevin Waugh, and Allan Schumacher. 2007. A Demonstration of SQUEGE: A CRPG Sub-Quest Generator. In *Proceedings of the Third AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Stanford, California) *(AIIDE'07)*. AAAI Press, 110–111.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. (2019). https://github.com/openai/gpt-2. Retrieved August 31, 2020.

[22] James Ryan, Ethan Seither, Michael Mateas, and Noah Wardrip-Fruin. 2016. Expressionist: An authoring tool for in-game text generation. In *International Conference on Interactive Digital Storytelling*. Springer, 221–233.

[23] James Owen Ryan, Casey Barackman, Nicholas Kontje, Taylor Owen-Milner, Marilyn A. Walker, Michael Mateas, and Noah Wardrip-Fruin. 2014. Combinatorial dialogue authoring. In *International Conference on Interactive Digital Storytelling*. Springer, 13–24.

[24] Judith van Stegeren and Mariët Theune. 2019. Narrative Generation in the Wild: Methods from NaNoGenMo. In *Proceedings of the Second Workshop on Storytelling*. Association for Computational Linguistics, Florence, Italy, 65–74. https://www.aclweb.org/anthology/W19-3407

[25] Judith van Stegeren and Mariët Theune. 2020. Fantastic Strings and Where to Find Them: The Quest for High-Quality Video Game Text Corpora. In *Proceedings of the 2020 Workshop on Intelligent Narrative Technologies*.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[27] Marilyn A. Walker, Ricky Grant, Jennifer Sawyer, Grace I. Lin, Noah Wardrip-Fruin, and Michael Buell. 2011. Perceived or not perceived: Film character models for expressive NLG. In *International Conference on Interactive Digital Storytelling*. Springer, 109–121.

[28] Nick Walton. 2019. *AI Dungeon*. Game [PC, Android, IOS]. https://www.aidungeon.io.

[29] Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-And-Write: Towards Better Automatic Storytelling. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.

[30] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 9054–9065. http://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf

[6] https://www.wowhead.com/news/pop-culture-references-in-wow-npcs-and-quests-from-games-and-music-228084