

Generating short RPG side-quest stories with Transformers

Jakub Myśliwiec
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
j.mysliwiec@student.utwente.nl

ABSTRACT

Progress in natural language processing has allowed for the exploration of automatically creating creatively written stories. The proposed research will exploit that possibility. By using GPT-2 trained additionally on a purpose build data-set the system will generate short stories to support RPG side-quests.

This paper's contribution is to show that fine-tuning a general pretrained transformer for story generation has potential to be a good option for obtaining descriptions for quests, if used in computer-human collaboration. It also highlights the shortcomings in the current method and proposes how those shortcomings may be addressed in future attempts.

Keywords

natural-language generation, Side-Quest generation, Computer Creativity, Transformers, Machine Learning, GPT-2

1. INTRODUCTION

An increasing number of tasks are being automated with the use of computers and this process is only getting faster. Even more interestingly, areas previously thought of as safe from automation, are being tackled. One such area is natural language generation. All thanks to a recent breakthrough: the Transformer [1]. It is a novel neural network architecture, that excels at natural language processing tasks. Another small but active area is narrative generation. There are many promising systems which can generate and store narratives in certain structures [2]. The efforts of generating surface text for those stories, so far, focused mainly on using rule based approaches [3]. It would be interesting to see how to use a Transformer to do the generation and how well that would work. Transformers have not proven themselves in writing narrative, so testing the limits of such models would be valuable. One area that lends itself well for this type of exploration is computer Role Playing Games and side-quests contained in them, as stated by Doran and Parberry [4].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

33th Twente Student Conference on IT July. 3rd, 2020, Enschede, The Netherlands.

Copyright 2020, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

Research Questions

- 1 Is it possible to train GPT-2 to generate a story suitable for a video game quest from just a title and an objective or summary?
- 2 How do the results compare to handwritten stories from existing games?
- 3 How could such a generator be used in practice?

The immediate aim of developing a quest generation system is to enrich the players experience when playing a Role Playing Game (RPG). In an RPG, a well written main story is usually accompanied by many side-quests. Commonly, they are non-obligatory small errands that exist to give the player more freedom [5]. Following a certain structure, they lend themselves well to automatic generation. My goal is to give them more depth by adding a short story.

First, a quest title and objective will be generated by some system, for the purpose of this research the system will be a human writing prompts manually. Second, a trained neural network is given the prompt and asked to generate a short story. Third, the system generates the story word by word. The prompt could take the form: *"Broken Reactor: Delivery 75 Electronic Parts"*. Then the generated text could be: *"I need you to run over to xxx and drop off these electronic parts. The chief technical operator there contacted me regarding the discovery of a major defect in their primary reactor. They don't have the parts required to repair it, so I want you to go and drop these off."* The example is taken from the Eve Info Mission database [6] and not generated.

A deeper goal is to test how capable of writing stories the Transformer architecture really is. So far, the possibility of generating deceptively human writing has been proven in certain domains. For example in writing propaganda more trustworthy than that written by humans [7].

Evaluating the outcomes qualitatively will allow for a comparison between the new method and traditional writing. Ultimately, when deployed, the output of this system is meant to be read by humans and convey certain information to them. Naturally then, asking human participants to evaluate the generated text validate the approach or prove that further work is required to reach usable levels of performance.

And finally, the achieved performance will be suggestive of taking a particular approach to generation as different levels of performance are best suited for varying ways of using them. A fully capable system could be deployed and directly interact with players. A slightly worse generator may be used to generate options for the game developer to choose from. One even less capable could be used as a brainstorming tool for writers or as an auto-completion

system. Such insights may also prove beneficial in determining a direction for future research in this area.

1.1 Justification

The application of this research could allow for improving the immersion in games where a player would be presented not just with a raw quest objective, but also a generated short story or a dialogue. Make the interaction feel more human through the addition of detail and character.

Moreover, games where each player gets a customized experience could become more complete and tangible with the inclusion of generated natural language that fills the world with little stories. The custom generated narrative would be done by a system like ReGEN [3] but a graph representation is not nearly enough to present the story in a compelling way. This is where my proposed generator would fill the gap and enable making such a game.

A previously mentioned benefit could be to greatly improve the productivity and the output of a writer. The system used as a tool that generates text quicker than a human would be able to write, it could greatly speed up the process or even free up time for the human to focus on other things.

When building on the proposed system there is potential to allow for more artistic control. Adding more parameters which can be tweaked could enable the artist to get more tailored text. Further, because language is a proven vehicle for stories, there is propensity for the output to be used as input for different systems that generate other forms of art which would attempt to tell the generated story. Examples could include music, video editing or even automatic creation of 2D/3D environments.

In the further parts of the paper, related work is referenced to show that no good system exists and requirements for an ideal system are defined. In the 3rd section I define and attempt to implement a generator to fulfill the posed requirements. The 4th part will be dedicated to verifying whether the defined requirements were met. Finally in the last section the research questions will be answered, conclusions drawn and a direction for further research proposed.

2. BACKGROUND

2.1 Related Work

A great deal of work has already been put into the research of narrative generation as stated by Kaybartas and Bidarra in their 2017 survey paper [2], even if talking specifically about quest generation. A summary written by Van Stegeren and Theune [8] gives a great overview of the different generative approaches that can be used to create stories. It does mention using deep machine learning, however in 2018 the main used model was the (now inferior) LSTM. More successful and notable approaches used context-free grammars, rewriting rules or predefined structures to generate the text. One of the early pattern and graph based approaches in the wider field is SQUEGE [5]. It relies on patterns it randomly selects and then populates the blanks with appropriate information about characters, locations and items. Authors of ReGEN [3] built on that idea and proposed a method for narrative graph rewriting, capable of creating complex branching stories. Their narrative building is impressive, unfortunately the conversion to text seems to fall to the defined rule-based approach. Creating a system to take graph representations and generate text would solve that problem however. The use of context-free grammars was also explored in RPG

quests [4]. All of the approaches relying on the mentioned structures face the problem of becoming repetitive.

What has significant positive impact of the feasibility of my research is the release of OpenAI's GPT-2 [9]; a General Pretrained Transformer in its second version. It is one of the largest and best performing general language models that is ready for more specific training. It is capable of generating arbitrary text which is largely grammatically correct and the content is coherent enough for some people to have difficulty identifying whether the text has been written by a computer or a human. Putting it to generating text that contains certain information was not one of the most likely tasks Radford et al. have foreseen however. Other insights into the use of Transformers like the contribution by Zellers et al. [7] help a great deal as well. They determined that a Transformer like GPT-2 can be given structured data to base the generated text on. It includes encapsulating it with special start and end tokens and using it as a prefix for the generation so that the Transformer can "pay attention" to it. Both transformers are capable of generating arbitrary text and are not shackled by the rule-based approach, however they are somewhat unlikely to add value to a player's experience out of the box. The texts generated are likely to be irrelevant.

There already has been a paper [10] which used Transformers with the idea of quest generation, though in this case, a quest could hardly be called narrative as it was designed to write cooking recipes. GPT-2 was trained to generate instructions based on the given title and ingredients of a recipe. The approach was found to be valid, though, to achieve statistically significant results, they would need to test their systems on more participants. Promisingly, the quality of the system with GPT-2 seemed to be roughly on par with human designed recipes in that particular setting. RecipeGPT [11] is another system designed around recipes.

2.2 Requirements for a generator

The knowledge gained by looking at related work allows for the identification of specific requirements for the generator.

The primary aim of including additional text with more information for the player is to enrich their experience. The text must add value for the player, otherwise there is no benefit at all in including it. Without added value text can at best clutter the game and at worst discourage the player from playing at all if they are forced to read it. Descriptions and dialogue must add value. This also means that it can not be too long (target median length is 4 sentences) as reading quests is usually not the main activity in games.

The proposed model should be able to generate anything. In the real world writers are not constrained to using specific patterns and structures in the texts they write. Sometimes the story requires text that does not strictly follow rules of grammar or even spelling. In contrast to more specific rule based approaches, the possible output space is much broader. A significantly larger proportion of the potential outputs then are meaningless and it is more difficult to get desirable results. However, it is appealing that there is fewer theoretical limitations due to inclusion of many more interesting possible outcomes.

A text narrative generation system will likely not be used only by tech-savvy people but also likely by writers without degrees in computer science. This is why it should be easy to use for people without technical knowledge.

Source	Title	Objective	Description
WoW	The Wayward Crone	Confront Helena Gentle in her home outside of Fallhaven	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
Destiny	Plaguelander	Complete Patrols in the Plaguelands	”We must fight to reclaim this land! From the Fallen, from SIVA. If we give an inch, we give the world.” -Shiro-4
EVE Online	Gone Berserk	Kill all NPC’s	I We have reports of Angel Cartel ships wrecking havoc in ***! Our convoys have been attacked, as well as some outlying outposts. They seem to be intent on causing as much damage as possible. I’d like you to take care of them, pronto! And be aware that they may not all be gathered together when you arrive and if so you will be able to pick them off as they warp in to the rendezvous point.

Table 1. Example quests from the dataset

Different users may also have different needs or preferences. Some may find a system that generates a ready to use text helpful. Others may want to write themselves but sometimes have a little bit of help. A system allowing for every degree of influence between the mentioned ones would be the most versatile.

The final requirement any system must fulfill is availability. If the target user is not able to access the system and get results, it could as well not exist at all. This is also the problem many outcomes of research face. After the paper is published, some people who read the paper would likely want to have the ability to run the application themselves. Unfortunately more often than not, even if the software is published together with the paper, it is not kept available as long as the paper is and to the detriment of advancement, building upon the work of predecessors is much more difficult.

3. METHODOLOGY AND APPROACH

The proposed system is a version of the GPT-2 model finetuned on a purpose-build dataset.

3.1 Gathering a Dataset

A number of non-trivial tasks need to be accomplished, the most difficult of which was the gathering of a dataset large enough to train GPT-2. Fortunately some games with active modding communities exist. Additionally many games have wiki pages which document the game. The main requirement was that the quests had to contain a description or dialogue and an objective which functioned as a summary of the description. After scouring the internet 3 websites with sufficient quality content were selected, all dedicated to popular MMORPG games: WOWHead [12] with World of Warcraft¹ (WoW) quests, TNR Destiny tracker database [13] for Destiny² and the EVE info [6] with EVE Online³ missions.

Later it was possible to harvest the necessary information by building a scraper which gathered it. The python library Scrapy⁴ was used to do the collection. It provides helpful utilities and allows the user to build an asynchronous ”spider” to parse html, follow links and extract information from predefined locations. Running the spiders to gather the information took a few hours in total

¹<https://worldofwarcraft.com/>

²<https://www.destinythegame.com/d1.html>

³<https://www.eveonline.com/>

⁴Scrapy|A Fast and Powerful Scraping and WebCrawling Framework <https://scrapy.org/>

due to the website’s restrictions on how often endpoints should be called not to overload the servers. Subsequently, the data could be inspected by drawing a random sample and manual inspection. If the texts prove to be of the quality and form expected from the end generator they could be merged into one large corpus to train on. Finally, after all data had been gathered it was cleaned to discard any duplicates and to remove any unusual characters or unwanted strings like placeholders or information about the quest being unused or deprecated.

Table 1 lists example quests taken from each dataset. The total number of collected examples from each source is shown in Table 2

Source	quests collected	after cleaning
WOWHead	24981	24872
TNR Destiny Tracker Database	680	595
EVE Infor Missions	310	310

Table 2. Summary of the number of quests collected per source

3.2 Training

Before GPT-2 could be finetuned several obstacles had to be overcome. First, the training data had to be converted into a format GPT-2 has been designed for. By default it includes a piece of text surrounded by special start and end tokens. Unfortunately, no way of passing labeled information has been foreseen or implemented out of the box. Luckily, Zellers et al. [7] came up with an approach that worked well on a neural network with similar architecture. They surround each piece of passed information with new tokens to help the transformer understand where the information is placed. The tokens can be anything as long as they are a short and unique string of characters. In the end each training example was constructed using the following equation.

$$result = " < |startoftext| > " + title + " < |obj| > " + objective + " < |text| > " + description + " < |endoftext| > "$$

GPT-2 was released in 4 different sizes, the larger, the better the performance. Large size also comes with a need for more data and the requirement of more capable hardware both to train and generate text. Due to the budget

Title	Objective	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 Gorggrond.	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.

Table 3. Generatd quest descriptions used in the survey

In addition to the 9 above also the first in Table 4 was used. The titles and objectives were selected randomly from the WoW dataset and given as a prompt for description generation.

constraints of this research the chosen platform to fine-tune the network on was Google Colaboratory⁵ thanks to free access to an instance with a data-science-oriented GPU. Not all the available GPUs support generation using the largest released model (1.5B) Moreover, training the model requires more resources and the most powerful GPU available (NVIDIA Tesla V100) is able to train only the the second largest model (774M). For those 2 reasons the 774M model was chosen. To train the network Max Woolf's gpt-2-simple⁶ was used. It is a python package that encapsulates the GPT-2 released by OpenAI. It was chosen because it provides simple access to the training

⁵<https://colab.research.google.com>

⁶M. Woolf. GitHub - minimaxir/gpt-2-simple: <https://github.com/minimaxir/gpt-2-simple>

and generation scripts in addition to useful utilities.

Several versions were finetuned with various training parameters and various compositions of the training dataset. Some were given less training steps (as low as 500) but a higher learning rate (5e-3 instead of 1e-5). The options were examined by generating up to 10 samples in order to determine what could be improved in the next training round. The final choice was between 2 models, one trained only on the data from WoW, the other on all all collected data. The former one was chosen as it offered the most consistency and it showed the most representation of details from the objective in the description. Upon further inspection of the datasets the ones not included proved to be flawed. The descriptions from the Destiny dataset too often included information unrelated to the objective.

The EVE Online dataset proved to have too little variation in the posed objectives. The same one: "Kill all NPC's" was repeated significantly too often. The training for the chosen model took approximately 4 hours.

3.3 Generation

Then with the model ready comes the final generation.

So how exactly does GPT-2 generate text? The process begins by converting the existing text to tokens and looking up the vector embedding for each token. This operation could be seen as locating a word in multidimensional space. Further, information about the position in text is added to each embedded token. Third, the a query vector is constructed from the last token. Fourth, the query token is used to determine which tokens from the input have the most influence on what will come next. Once again, the analogy of another multidimensional space can be used. In this case the importance of a token is determined by looking at the distance to the query point. The closer, the more important. Fifth, a vector, which can be compared to a weighted average, is computed from the all previous tokens, weighted on importance. It can be seen as a point in the first mentioned space. Sixth, a set of closest embedded tokens are selected and given a probability. Finally, one is chosen to be the next generated token. This way the next token has been generated and the process can start from the beginning, this time with input being longer and containing the newly generated token. The process is repeated until the model decides to generate a special `<|endoftext|>` token.

The most easily influenced step at this stage is the sixth and seventh step as described in the preceding paragraph. Three parameters can be given to influence how many words the system can choose from by using (top_k to cut off how many can be in the set, top_p a maximum "distance" the point can have to still be included in the set)

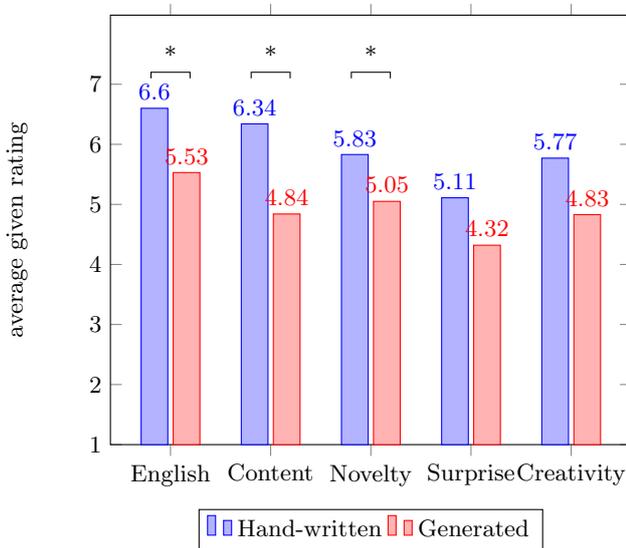


Figure 1. Comparison of average ratings on specified criteria

Scores collected from n=32 participants in a survey where 10 examples quests of each source were rated on a Likert scale 1-7. Criteria: English - spelling and grammatical correctness,

Content - information content about the goal, Novelty,

Surprise, Creativity - participants perceptions on the text having the qualities designed to measure value. The * signifies statistical significance with $p < 0.01$ for English, Content and Novelty.

and how likely the system is to generate less expected results (temperature influences how unpredictable the results are).

To pass the title and the objective, so a description can be generated for them, the beginning part of the training structure is used. The prefix used to generate description examples seen in Table 4:

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

Interestingly, the model has learned the structure so well that it can often generate entire quests by itself. They follow the exact same formula with parts split by tokens and also seem to be usable. Because all examples the model knows were taken from WoW, all the generated quests mostly follow the lore of the game world.

When using google colab the generation of a single quest takes around 40 seconds. However it has been optimized to generate in batches and obtaining 5 descriptions for the same prompt takes only 60 seconds. Initial setup adds about 6-7 minutes to the start of the process, but needs to be done only once per session.

4. VERIFICATION

4.1 Experiment Design

To perform quantitative comparison of the created generator an online survey was conducted. The participants were presented with 20 examples of quests which consisted of a title, an objective and a description. Ten of the quests were randomly chosen from the training set, for the other 10 a title and a description was randomly selected from the dataset and the algorithm was asked to generate the description. All were generated with temperature=0.9. Finally all 20 were randomly ordered. For each quest the participants were asked to rate on a Likert scale (1-7) how much they agree with 5 statements. Each one was designed to measure a separate property.

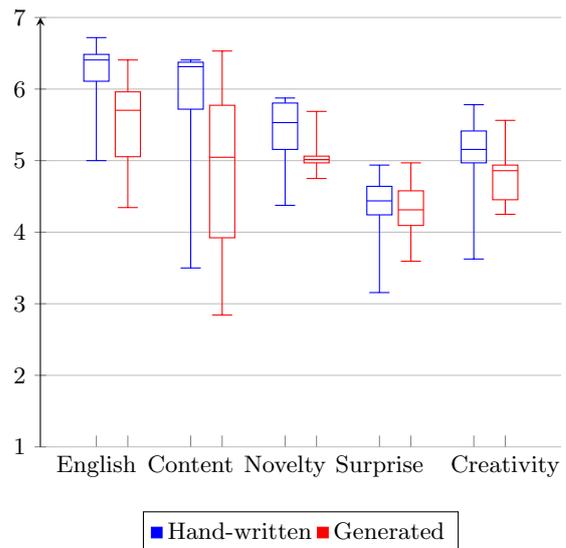


Figure 2. Distribution of scores per question

The high variability in obtained scores suggests that the generator is capable of producing high quality output. It could be harnessed by manually selecting the best example from the generated.

1. English - "The description makes use of correct English"
2. Content - "The goal is clear from the description"
3. Novelty - "The description is written in a novel way"
4. Surprise - "The description is surprising"
5. Creativity - "The description is creative"

The last three properties were selected specifically to measure creativity based on Boden's breakdown [14]. The author stipulated that creativity can be measured in terms of novelty, surprise and value. Based on preliminary testing of the survey questions it was decided to substitute value with creativity itself as participants found the question about value too ambiguous and meaningless. The survey design was inspired by the one conducted by Ammanabrolu et al. [10].

In addition to the main part, two more questions were asked. Everybody had to state their level of English knowledge on the CEFR scale [15] and their previous level of familiarity with RPG computer games on a scale from 1 to 5 in case additional analysis based on those criteria brought insights into scores that differ from the rest. The analysis of the differences however brought no interesting information.

4.2 Results

Before comparing the results from the survey, inspecting the example results of generation leads to worthwhile findings. Just purely reading the examples can show the remarkable quality of the structure, grammar and spelling. This is in part thanks to the creators of GPT-2 and the training they did. It really has learned the basics of English. Moreover the finetuned example has also learned that it is good to mention something from either the title or the objective in the generated description. In the first and last example a name from the objective is mentioned. This would not have been achievable with technology (LSTM) used prior to the Transformer. Unfortunately the decision to generate examples using temperature=0.9 makes the output somewhat unpredictable. In comparison to more traditional methods of text generation, this offers much less control and likely more variety.

In the end n=32 participants filled in the survey. First, for each participant the scores were grouped into 2 categories: generated and handwritten. Averages for each criterion were calculated and for those scores the sign test on the median was conducted as none of the differences showed a normal distribution. The resulting scores are visible in Figure 1. Three parameters: English, Content and Novelty were shown to be statistically significantly worse in descriptions generated by the model with $p < 0.01$. The two other statistics: Surprise and Creativity however promisingly did not show significant differences even with

Temperature = 0.9

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.

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.

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.

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.

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

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.

Master Ashwall wants you to head to the Purple Parlor in Zoram'gar Outpost and inform them of what you have learned.

Hear out the Council of Six in the Purple Parlor!

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 (arguably too low)

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.

Table 4. Examples generated from the same prompt by finetuned GPT-2

The first 4 examples were generated with temperature = 0.9. The latter 4 were generated with temperature=0.7. Notice the wide range of quality between texts generated with the same temperature. Additionally, it is interesting that the system generated a story which is exactly the objective contained in the prompt in the 3rd of the 0.7 examples. Moreover, notice that the ones generated with higher temperature contain more unorthodox spelling and are more unexpected.

The last example was selected to demonstrate the drawbacks of a too low temperature.

$p < 0.05$.

These results are promising as the generator does not fall behind on every aspect. Moreover, the statistically significant differences are not large either. The fact that Surprise and Creativity are the closer ones may be influenced by the high temperature setting when generating the texts. Similarly the differences in language quality and content may also be increased by high temperature. Though, there still is a possibility that the good results speak more to the detriment of WoW quests quality rather than high quality of generated text.

Upon closer inspection of the rating distribution across the questions, more valuable information can be gained. A summary can be seen on Figure 2. In most cases it can be seen that the generator is capable of producing results that are above the median handwritten texts. In fact sometimes quality even exceeds that of handwritten examples in the Content and Surprise criteria. The inconsistency of good results suggests taking a cherry-picking approach to generation. Instead of generating one description the system can be used to generate multiple and the user can inspect and choose one that best suits the needs. Multiple descriptions generated from the same prompt can be seen in on Table 4.

If the desire is to use the results of generation without any additional inspection then generating the texts using a temperature=0.7 or even lower may prove beneficial. As seen in Table 4 texts generated at a lower temperature seem to be more consistent both in spelling the the content of information. Generating at lower temperatures grows more predictable, however texts generated with temperatures that are too low get highly repetitive.

5. CONCLUSION

Generating quest descriptions using a finetuned version of GPT-2 is a viable approach. Without additional work the generated text is slightly worse than handwritten examples but much cheaper. Putting in some time and attention can likely greatly improve the quality of the obtained results and may lead to texts good enough to be used in a product without hesitation.

The approach of passing arguments to GPT-2 by delimiting them with unique tokens proved to be a success. GPT-2 learned to use the passed information to inspire the text it generates, though, the results may be slightly inconsistent.

Generated texts measure up beneficially to handwritten texts when looking at the criteria of surprise and creativity. Without human supervision, the quality of English and the content of information relevant to the objective are demonstrably sub-par. However, by generating multiple examples and choosing the best or by generating texts at a lower temperature the results can be improved to reach satisfactory quality.

The created model can be used for all mentioned use cases with varying levels of success. The best case would be to use it as a tool for a writer who either reads the generated text as inspiration, incorporates it into their writing or even simply chooses the best generated option. More automatic approaches are also possible, though to a detriment in quality which must be evaluated on per use case basis.

To improve the quality of English in the generated texts training the larger (1.5B) GPT-2 model is likely to be beneficial. Additionally, more quantitative analysis con-

cerning parameter adjustment and the (un)supervised approach to generation should be further tested to explore the theories drawn from qualitative analysis performed during this study. Further improvements to the system include building a more diverse dataset of training examples and considering adding more parameters controllable by the user such as the genre (fantasy/sci-fi).

Finally I have decided to share the results of my work so that you can test the system and see how it performs on your own examples.

<https://jakub.thebias.nl/research/QuestGen/colab/>

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, I Guyon, U. V. Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, Eds., Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.
- [2] B. Kybartas and R. Bidarra, "A Survey on Story Generation Techniques for Authoring Computational Narratives," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 9, no. 3, pp. 239–253, Sep. 2017, ISSN: 1943068X. DOI: 10.1109/TCIAIG.2016.2546063.
- [3] B. Kybartas and C. Verbrugge, "Analysis of ReGEN as a graph-rewriting system for quest generation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 6, no. 2, pp. 228–242, 2014, ISSN: 1943068X. DOI: 10.1109/TCIAIG.2013.2290088.
- [4] J. Doran and I. Parberry, "A prototype quest generator based on a structural analysis of quests from four MMORPGs," in *ACM International Conference Proceeding Series*, 2011, ISBN: 9781450308724. DOI: 10.1145/2000919.2000920.
- [5] C. Onuczko, D. Szafron, J. Schaeffer, M. Cutumisu, J. Siegel, K. Waugh, and A. Schumacher, "A Demonstration of SQUEGE: a CRPG Sub-Quest Generator," Tech. Rep., 2007.
- [6] *EVEInfo / EVE Online Missions*. [Online]. Available: <https://eveinfo.com/missions/> (visited on 05/07/2020).
- [7] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending Against Neural Fake News," May 2019. [Online]. Available: <http://arxiv.org/abs/1905.12616>.
- [8] J. van Stegeren and M. Theune, "Narrative Generation in the Wild: Methods from NaNoGenMo," in *Proceedings of the Second Workshop on Storytelling*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 65–74. DOI: 10.18653/v1/W19-3407. [Online]. Available: <https://www.aclweb.org/anthology/W19-3407>.
- [9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [10] P. Ammanabrolu, W. Broniec, A. Mueller, J. Paul, and M. O. Riedl, "Toward Automated Quest Generation in Text-Adventure Games," Sep. 2019. [Online]. Available: <http://arxiv.org/abs/1909.06283>.

- [11] H. H. Lee, K. Shu, P. Achananuparp, P. K. Prasetyo, Y. Liu, E.-P. Lim, and L. R. Varshney, "RecipeGPT: Generative Pre-Training Based Cooking Recipe Generation and Evaluation System," in *Companion Proceedings of the Web Conference 2020*, ser. WWW '20, New York, NY, USA: Association for Computing Machinery, 2020, 181–184, ISBN: 9781450370240. DOI: 10.1145/3366424.3383536. [Online]. Available: <https://doi.org/10.1145/3366424.3383536>.
- [12] *Quests - World of Warcraft*. [Online]. Available: <https://www.wowhead.com/quests> (visited on 05/03/2020).
- [13] *Quests - Destiny Tracker Database*. [Online]. Available: <https://db.destinytracker.com/d1/quests> (visited on 05/13/2020).
- [14] M. A. Boden, "Creativity in a nutshell," *Think*, vol. 5, no. 15, pp. 83–96, 2007, ISSN: 1477-1756. DOI: 10.1017/s147717560000230x.
- [15] *International language standards | Cambridge English*. [Online]. Available: <https://www.cambridgeenglish.org/exams-and-tests/cefr/> (visited on 06/19/2020).