Dungeons and Datasets: Final Project Report

Zach Grow, Joel Williams and Andrew Ruskamp-White

Portland State University

Abstract

Our aim for this project was to create a tool for use during sessions of Dungeons & Dragons or other similar high-fantasy tabletop role-playing games (TTRPGs). We also wanted to explore the practical obstacles and methods involved in dataset creation, as there is not a lot of prior documentation on that subject. We decided to make a dataset of dungeon room descriptions, as there are plenty of tools in this space already but none that make use of NLP techniques. We used a combination of human curation and algorithmic scoring to build out the dataset. Our final dataset contains over 2000 distinct sentences. Though small, this dataset proves big enough to generate novel results on natural language generation tasks. [1] We also consider the legal and ethical implications of building datasets from the work of artists and authors, even for a Free and Open Source project.

1 Introduction

Tabletop role-playing games such as Dungeons & Dragons ©(shortened D&D) have a close link with natural language, in that a "successful" game of D&D requires participation in both a performative mode, as experienced during a "live" game session, and a descriptive mode, experienced through the consumption of written text produced to support or inspired by some particular game. The leader of a group of players, known as the "Dungeon Master" (DM), is responsible for the bulk of the creative work required for a successful game. This creative work must be thematically consistent and interesting, to hold the attention of the players, but is also often repetitive and comes with no guarantee of use; there is every possibility that the DM may generate descriptions for rooms that the players never go into. Many DMs fall back on automated tools to offset this labor cost, saving their creative energy for more important and more visible parts of the game. In this paper, we investigate the current possibilities of using natural language processing tools for the generation of room descriptions within the larger context of tabletop role-playing games, such as D&D. We provide a summary of current work, and describe our own efforts to extend this work. We outline the work performed in creating a dataset for the specific purpose of generating room descriptions, including legal considerations and development of a prototype generative model that uses ML and NLP techniques as opposed to combinatorial methods.

We began by looking for other datasets suitable for NLP tasks. We found only one, the Critical Role Dungeons and Dragons Dataset (CRD3), which consists of the set of transcripts of all episodes of the Critical Role live-stream show. The Critical Role show is a "live" broadcast of a D&D game as it would be run at home; therefore, the transcripts provide a snapshot of how the game is experienced as a group performance, including conversational asides, meta-game conversation, cultural references, and so on. This dataset was originally created as a tool for abstractive summarization tasks; however, we were able to modify the dataset for our purposes by the expedient of removing everything that had nothing to do with a room description.

The resulting dataset was small: we therefore investigated a number of ways to extend the dataset. We hoped to obtain samples from unique and original contexts, such as from other transcripts, but ran into legal and time constraints. We therefore supplemented the dataset by adding results from a "traditional" combinatorial algorithm and by human curation of freely available published material, which was large enough to enable the development of a generative model. We went on to enrich the dataset by taking examples from published sources,

[1] All code and the dataset can be found at: https://github.com/arusk2/dungeons-and-datasets

with the awareness that the resulting dataset would be therefore ineligible for distribution.

We built a simple generative model as a proof-of-concept and to obtain some feedback on whether our dataset would produce meaningful results. The dataset was preprocessed for invalid data and duplicates, and then passed through a pre-trained roBERTa model. The results were surprising, as we will discuss in the Results section. Despite the small size of our initial dataset, and the simplicity of our model, the output was promising, though inconsistent in quality and coherency. A larger table with examples can be seen in Appendix A.

We go on to suggest some strategies and concepts for creating and extending both a distributable dataset of room descriptions and a generative model for same. Our methods for extracting room descriptions from transcripts of "live" play could be applied to transcripts of "play-by-post" games, which are conducted asynchronously via posts on a community forum. There are a number of legal considerations specifically relevant to the creation of this dataset; we outline some solutions for satisfying these considerations while still maximizing the natural origins of the data itself. We also describe ways in which the generative model could be extended, both in terms of feature additions and improvements to the generation method itself.

2 Related Work

Generative models for specific tasks are fairly commonplace, however a specific room and location description dataset for NLP language models is relatively unheard of in the academic literature. The Dungeon Master's Guide itself has a rollable dungeon generation table, but the results are usually unrealistic and frequently repetitive which breaks player verisimilitude. Online generators have proliferated on the internet, but we have not encountered one yet that uses a robust NLP model. The ones we have seen are based on random combinatorial methods, essentially a glorified home-made digital version of the tables available in the Dungeon Master's Guide.

That being said, there is some academic research that has been centered around Dungeons & Dragons especially in regards to general NLP concepts. In 2019, MacInnes published *The D&D Sorting Hat: Predicting Dungeons and Dragons Characters from Textual Backstories* (MacInnes, 2019) which was about generating D&D character statistics from their textual backstories using a variety of machine learning methods. We found this work to be extremely compelling, however perpendicular to what we wanted to accomplish.

Another study published in May 2021 by Man Si et al, entitled *Telling Stories through Multi-User Dialogue by Modeling Character Relations* (Si et al., 2021) used the Critical Role Dungeons and Dragons (CRD3) dataset (the same dataset we used for our project, which contained unfiltered dialogue of D&D play) to model relationships between various characters.

3 Methodology

3.1 Human Curation

Our initial drive was to use a web scraper to crawl data from forums where users play TTRPGs by posting on public forums. However, we quickly realized we did not have a dataset that we could train a model on that could automatically identify relevant room descriptions from the variety of other table talk, like character interactions, combat, or cross-talk and jokes between players.

There are also legal issues that we encountered regarding automatic web scraping that we discuss in Ethical Considerations. To mitigate this, we decided to make our first move to collect data from published, pre-written adventure modules available for free on `www.dmsguild.com` and `www.drivethrurpg.com`. We also found and used several random room description generator websites (Ball and Ball, 2019a,b,c; Williams, 2016; Harlan, 2018; Stronski, 2021; Kositz, 2021, 2019; Doty, 2022; Simoes, 2022; Club, 2019; Gubitosi, 2022; Winegar, 2016; Generators, 2022; Net, 2021; Perchance, 2021; Herridge, 2020; Planner, 2021).

Samples of data from these sources were collected by each of us manually and selected based on criteria established by the research group, namely: is the selected text used to modify the appearance of a room being described? This criteria was left intentionally broad to allow for the selection of statements ranging from "The room is 10 feet wide by 20 feet deep and has a bare floor" to "Crystals are arranged in a circle across the table." The variety in selection helps us build a dataset that can be used to create many unique room descriptions when used to fine-tune a natural language generation model.

Seed phrase	Generated Text
"You enter"	You enter the room with a rough impression of what might have been once an abandoned bodega.
	You enter a small room filled with a bowl of fruit and vegetables in the center of the room.
	You enter a room dominated by the stench of moldy stench.
"The room"	The room contains a variety book, the next coming will be: "The Golden Age".
	The room is a collection of bookshelves, texts, and even a large library containing many texts.
	The room smells of cooking, which is somewhat like cooking fumes coming out of the oven before you open fire.
"The door"	The door opens without much difficulty and the party waits until they're able to enter the room.
	The door behind you slowly opens in silence to reveal the sight of a large reptilian humanoid.
	The door seems to be unlocked and locked.

Table 1: Examples generated from a GPT-2 model fine-tuned on our dataset.

3.2 Algorithmic Curation

The act of collecting our data could not be solely done by hand. The amount of text to process would not be feasible to evaluate in reasonable time. So, after an initial manual collection of just over 2000 sentences, we created a brute force algorithm for scoring future data and extracting room descriptions. This algorithm can be used on any text based corpus to augment our dataset. For this project, we used the algorithm on the Critical Role Dungeons and Dragons Dataset (CRD3) (Rameshkumar and Bailey, 2020).

We preprocessed the data to remove duplicate lines. Since entries were in the form of transcript episodes from the podcast Critical Role, introductions and closing statements were frequently repeated and were removed. Then we used a pre-trained roBERTa transformer model from the spaCy natural language processing library to obtain a tokenized, parts-of-speech-tagged and lemmatized version of the CRD3 dataset. We then score each line based on criteria that will be discussed in the next section. Positive scoring sentences are then human reviewed for selection in our dataset.

3.3 Scoring Sentences

First, we defined a set of positive scoring criteria based on parts of speech identifiers within a certain sentence structure[2], vector-based similarity with commonly occurring words that we extracted from our hand-crafted dataset, and whether the sentence started with "You".

Criteria	Score
Sentence begins with "You" lemma	+1
Comparing a stored list of parts of speech tags with examples	+1
Vector similarity > 0.5 "room"	+2
Vector similarity > 0.5 "floor"	+2
Vector similarity > 0.5 "door"	+1
Vector similarity > 0.5 "wall"	+1

Next, we defined a set of negative scoring criteria based on token similarity, the presence of tokens that implied dialogue, the presence of interjections, curse words, and named entities.

Criteria	Score
Sentence begins or ends with " token	-1
Sentence ends with ? token	-1
Vector similarity > 0.5 with curse words	-2
A token mid-sentence match with "	-2
A token tagged with interjection part of speech	-2
Presence of tagged named entities	$-2 \cdot n$

We found after experimentation that vector similarity greater than 0.5 covered a reasonably wide swath of vectors related to the word in question and rejected enough outliers to be extremely effective. After adding other word vectors to compare

[2]This method is better elaborated by looking at our repository code and we admit that our method is rather primitive.

to that were also high in occurrence in our handcrafted dataset, we saw a blanket increase in false positives in location descriptions and not a noticeable increase in caught examples. We found that the positive scoring criteria, which often double counted various traits we found desirable, along with negatively scoring undesirable traits, cut our CRD3 dataset down by at least 93% from its raw form and presented us with, although not exactly perfectly delineated, excellent location description examples when we excluded all scores less than 1. Many of the best examples scored over 10 points.

4 Experiments

Because our project was centered around generating a dataset, we didn't focus on performing experimentation, although we did investigations with a rudimentary generative model, simply to determine whether our dataset was going to be feasibly useful. We used TensorFlow to train a GPT-2 model from Hugging Face that used an Adam optimizer. Initially, when our dataset was particularly small, results were expectantly poor. Generative models starting with "You enter" were nonsensical and repetitive. As we grew the dataset and fine-tuned the parameters of the model, the results improved.

Table 2 displays some results of a model trained on our dataset. We fine-tuned the model for only 2 epochs, so compute time to generate these examples was only about 15 minutes.

Appendix A displays some of the results of the same model trained on our final alpha release dataset. We trained the model for 5 epochs, since we noticed a slower convergence with a larger dataset. Compute time was similar to the first early examples.

5 Results

Our alpha release of our dataset contains 2,000 example sentences of room and location descriptions in a high-fantasy world. The data was curated and handpicked by us, and while there are some short and not as high quality examples, after fine-tuning our rudimentary test generation model we got interesting and expressive room descriptions that we generally considered to be on par with other online room description generators that work entirely by random combination. We did not have a qualitative measurement for evaluating machine generated sentences with human generated sentences, but work in this space could be exciting. With further work improving the generation model and adding more examples for a beta release, we believe that we can eventually create a room generation model that far exceeds the quality of currently available online random generators in terms of its expressiveness, realism, and capacity for generating novel descriptions.

We are open to expanding this dataset more, though this will require more work and cooperation with forums or authors to create more examples for the dataset. We plan on continuing this project and compiling a beta release.

6 Conclusion

Given more time, we could have produced a dataset that was up to our standards and preferences for quality and originality. The larger obstacles we faced were due to our inexperience with the process, and with the legal obstacles related to the source material, as opposed to technical limitations. Dataset creation in general is quite feasible, even straightforward, but the sheer size of the average dataset still requires some creative problem-solving. We have a few advantages in this field that make the production of this dataset still feasible given enough time: the existing culture of home-made content within the D&D community; the proximity of FOSS paradigms to our project space; and the pursuit of creative goals as opposed to strictly research or analysis.

First, there is a strong tradition of players and fans of TTRPGs building, testing, and sharing their own content with others in the community dating back to before personal computing. By participating in this, and in the gift community at-large, we hope to rely on goodwill and interest where we can't rely on a strong base of capital.

Second, there is a high correlation between software programmers and the TTRPG community. This correlation has caused precedent to be established in this field with respect to creative licensing. [3] We can borrow from their solutions to ensure equitable treatment for everyone involved.

Third, our creative goals mean that some of the parameters we wish to optimize will naturally come about as our dataset and generative models improve. Preliminary testing was optimistic about the impact of naturally- vs computer-generated descriptions

[3] For example, Wizards of the Coast ©has published a stripped-down version of the D&D core rules under a fair-use license since 3rd Edition, in 2000.

on the dataset. Each new author who contributes to our database means an additional sample of variance, syntax, and style, and each new description increases the domain of language that the generative model has access to. This can help fine-tune a more robust and expressive natural language generation model. Unlike a statistical model, we want to optimize for novelty, suggestion, and creativity, all in the name of more fun.

7 Interesting Insights

None of us had compiled a dataset before, so we encountered several spiked pit traps that we did not foresee. For those interested in compiling their own datasets or interested in contributing to ours, we offer these points of consideration before embarking:

- Consider the legal/ethical considerations heavily before you begin. We believe authors should be paid for their work, licences, and TOS agreements respected.
- Start early and consider the amount of time it takes simply to find good examples. We spent hours just gathering and formatting data.
- Spend time early considering how exactly you're going to source your examples. We had to scrap our whole first plan to scrape after encountering difficulties.
- Even simple parsing and tagging of a giant dataset like CRD3 (with millions of example sentences), takes a very long time. Running our spaCy NLP pipeline takes over 8 hours on a normal computer.
- Our rudimentary scoring algorithm for the CRD3 dataset worked surprisingly well for our purposes. Sometimes doing something in a rudimentary way instead of "the perfect way" is the best path.

8 Ethical Considerations

The worlds of independent publishing and homebrew content production for TTRPGs have always been tightly linked, and therefore sometimes contentious. We wanted to be compliant with FOSS, so as to maximize the accessibility of our tool, but our dataset's contents are unlike other NLP datasets in that the data itself is a form of creative performance and therefore subject to intellectual property laws. This seems roughly analogous to training an image generation model on images whose copyright belongs to someone else. In addition, the process of scraping a forum can often be stressful on the servers that host it. The terms of use generally forbid this without prior authorization. We queried some of the forums management anyway to see if they might make an exception on educational grounds, but were denied in all cases. To overcome these obstacles, we propose the collection of room description samples as an open call for donations, similar in process to how publishing companies solicit submissions for new works. Our results demonstrated that even a small amount of "rich" text provides a strong improvement in the quality of the model's output; by crowdsourcing the dataset, we can maximize the overall originality of the data. This also offers the advantage of establishing a clear legal position with respect to the creative input and output of the project.

9 Future Directions

Due to the limitations on our ability to scrape "play-by-post" forums, we thought it would be a good idea to instead do an open call for DMs and players to submit their own original work. Normally, putting an open call out to the internet community for dataset material receives a lukewarm response, but considering the high level of fandom of D&D and our plan to eventually open source this future dataset for that same community to enjoy, we thought it would be worthwhile to try. Furthermore, our dataset has examples which were sourced from random generators, which we do not consider to be an ideal source as format and tone rarely vary. In order to curate an expressive and maximally humanistic dataset, we really wanted to have entirely human written examples. Due to legal and time constraints, we were forced to resort to using some random generation. A future ideal dataset would be willingly gathered and maximally human, avoiding the need to negotiate creative licensing issues that arise from forum scrapes and using PDF material. Finally, since both our generative model and room/location scoring algorithm were relatively primitive, we would also like to put more work into improving both. Training a transformer model to recognize room descriptions (something of a dragon before the dragon egg problem), as well as incorporating word2vec or similar tools into the generative model are future goals.

10 Acknowledgments

The authors would like to thank Gary Gygax, the creator of Dungeons & Dragons; Wizards of the Coast©, the intellectual property right owners of Dungeons and Dragons©; The Portland Game Store, Guardian Games, and all of the other local game stores in Portland that didn't make it through the pandemic. Support your local game store! We would also like to thank all authors and contributors from sources we used in the References section, whom are numerous and have written fantastic adventures.

References

Jonathan Ball and Beth Ball. 2019a. *First Blush*. Dungeon Master's Guild.

Jonathan Ball and Beth Ball. 2019b. *Second Glance*. Dungeon Master's Guild.

Jonathan Ball and Beth Ball. 2019c. *Third Time's the Charm*. Dungeon Master's Guild.

Quickphix Club. 2019. *Bound in Chains*. Dungeon Master's Guild.

Justin Doty. 2022. *The Summoning*. Dungeon Master's Guild.

Fantasy Name Generators. 2022. Fantasy Name Generators dungeon description generator.

Richard Gubitosi. 2022. *A Rat Problem*. Dungeon Master's Guild.

Dan Harlan. 2018. *Dungeon Room Descriptions*. Dungeon Master's Guild.

Daniel Herridge. 2020. *Crypt of the Archmage*. Dungeon Master's Guild.

Bryce Kositz. 2019. *Reliquary*. Dungeon Master's Guild.

Bryce Kositz. 2021. *Temple of the Nightwind*. Dungeon Master's Guild.

Joseph C MacInnes. 2019. The d&d sorting hat: Predicting dungeons and dragons characters from textual backstories.

Roll 4 Net. 2021. Dungeon rooms.

Perchance. 2021. Perchance room generator.

DND Campaign Planner. 2021. Room description generator.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A critical role dungeons and dragons dataset. Association for Computational Linguistics.

Wai Man Si, Prithviraj Ammanabrolu, and Mark O Riedl. 2021. Telling stories through multi-user dialogue by modeling character relations. *arXiv preprint arXiv:2105.15054*.

David Simoes. 2022. *Deicide*. Dungeon Master's Guild.

Benedikt Stronski. 2021. *Temple of the Nightwind*. Dungeon Master's Guild.

Steven Williams. 2016. *Nerzugal's Dungeon Master Toolkit*. Dungeon Master's Guild.

Michael J. Winegar. 2016. *Leomunds Misplaced Manor*. Dungeon Master's Guild.

A Expanded Generated Examples

Below is an expanded list of examples generated from a GPT-2 model fine-tuned on our dataset.

Seed phrase	Generated Text
"You enter"	You enter a narrow hallway with dusty furniture, a dusty storage, and a few other artifacts scattered around.
	You enter a cavernous room, containing another half dozen rooms of this kind.
	You enter a large, dusty, worn-down hall.
	You enter a darkened room with a faint but audible scream from the darkness behind.
	You enter a small room, with a single window.
	You enter an undisturbed chamber filled with an eerie, metallic effect.
	You enter this chamber, and enter a small, dim room with flickering torches that are in disarray.
	You enter this square room.
	You enter a temple. 40 feet down you find yourself near the center of this temple.
	You enter a large cavern, with a giant wooden door at the end of the wall.
"The room"	The room has four windows, two on each side of the room, and two on each side of the door.
	The room has a carved ceiling that has a carved-wood floor.
	The room is decorated with a golden robe's sconcese on its chest and a golden plate is mounted on an arm's staff decorated with intricate runes.
	The room is covered with a dusty floor.
	The room contains a locked door and a trapdoor on the north-east wall.
	The room is full of large bookshelves filled with exotic flora and fauna, and a shelf with a very comfortable leather armorial.
	The room is not entirely dark, but some dim vision patterns emerge.
	The room is divided roughly into two, with a few small, but sturdy chairs.
	The room smells of mold and mould.
	The room smells of decay.
"The door"	The door to this chamber opens into a dark hallway with small metallic tubes that lead right through this chamber.
	The door is made of pine wood, and the inscription reads: The priests's house is to be destroyed.
	The door appears to lead you through a small trapdoor in the bottom of a small room that has an iron bolt inside it, a door that appears to lead you on the opposite side of the room while the trap is in place on the door and so on.
	The door lies silent at the moment, with a faint creak.
	The door leads into a basement containing what looks to be a large laboratory.
	The door shows signs of being unlocked.
	The door handle is made of a simple stone and hinges upon one side of the door.
	The door has a large square bolt in its center as well as a pair of small bolts on the side as well as a set of glowing orbs on the far side.
	The door to the north is made of iron, and the inside is a bit rough and cracked.
	The door opens to reveal another room without a word for a moment.

Table 2: Examples generated from a GPT-2 model fine-tuned on our dataset.