Natural Language Generation With Fine-tuned GPT-2 Models

Cole G. Gilmore
Boise State University
Boise, Idaho
colegilmore@u.boisestate.edu

ABSTRACT

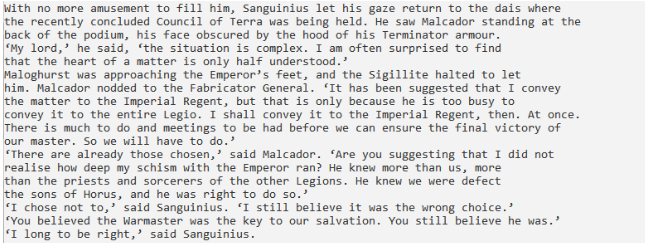
In today's world we are seeing a rapid increase in the world of AI generated content, especially with AI generated text documents. As such, it is a relevant area of study to determine just how easy and accurate a natural language generator can produce output of a certain genre. The direction taken was to try and produce models that could accurately generate a small chapter of text from the popular Warhammer 40K novels. Therefore, the general design of this project was to structure data into a special format for testing and then using cross-validation loss in conjunction with multiple libraries to determine accuracy. As well as getting user responses to help determine how convincing the sample text output was. In the end it was determined that over training was a genuine concern and it became difficult to determine exactly when to halt training of the model(s). Another issue was being able to train the larger GPT-2 models as they are incredibly computationally expensive and slow to train on a single machine. However, with some tweaking of existing models, very accurate and convincing samples of text were able to be generated in multiple different formats.

1 INTRODUCTION

Being able to discern human generated content from artificially generated content is becoming vital to browse the internet in current years. As, who is to say that the news article you got a notification to read this morning was really written by a person, and not just a machine told to write a story given only a handful of facts? For this reason, it's important to get a good background on just how this can impact regular internet users as it's commonly said that artificial intelligence algorithms will become commonplace in generating news articles in the next ten to twenty years.

Therefore, it's not too far out of the realm to say that artificial intelligence algorithms will eventually be writing all of the news articles. Whether that is in twenty years or forty, it's likely to happen at some point in our lifetimes just with the current direction natural language generation is taking. Especially with new language models being released fairly regularly, with a recent example being the release of an early version of GPT-3, we are quickly approaching a point where artificially generated content is indistinguishable from human generated content unless it is examined with a fine tooth comb.

With that in mind, I determined that my goal was to be able to generate a completely random small chapter sample of text. Then, see just how accurate it was compared to the original data and whether or not it was able to fool someone into thinking the sample was taken straight from a human written book. With the final step being whether or not the generated sample of text is able to fool another person very familiar with the novels that the data was gathered from. With some example samples being fairly close to fooling all parties. Figure 1.



```
With no more amusement to fill him, Sanguinius let his gaze return to the dais where
the recently concluded Council of Terra was being held. He saw Malcador standing at the
back of the podium, his face obscured by the hood of his Terminator armour.
‘My lord,’ he said, ‘the situation is complex. I am often surprised to find
that the heart of a matter is only half understood.’
Maloghurst was approaching the Emperor’s feet, and the Sigillite halted to let
him. Malcador nodded to the Fabricator General. ‘It has been suggested that I convey
the matter to the Imperial Regent, but that is only because he is too busy to
convey it to the entire Legio. I shall convey it to the Imperial Regent, then. At once.
There is much to do and meetings to be had before we can ensure the final victory of
our master. So we will have to do.’
‘There are already those chosen,’ said Malcador. ‘Are you suggesting that I did not
realise how deep my schism with the Emperor ran? He knew more than us, more
than the priests and sorcerers of the other Legions. He knew we were defect
the sons of Horus, and he was right to do so.’
‘I chose not to,’ said Sanguinius. ‘I still believe it was the wrong choice.’
‘You believed the Warmaster was the key to our salvation. You still believe he was.’
‘I long to be right,’ said Sanguinius.
```

Figure 1: A small generated sample of text, using the GPT-2 355M model as a base.

2 CORE STRUCTURAL ELEMENTS

2.1 Background

For the vast majority of this project the GPT-2 models provided by the openai team, developers of the GPT models, were used as they are mostly pre-trained and contained most of the work required to generate random samples of text. As such, there weren't many classes that I had to make to begin training the models from GPT and calculating validation loss. With the initial direction of the project being to create an interactive text generating game, similar to a choose your own adventure book, much of the initial work was based off of the open-source AIDungeon project. However, due to the direction switch of the project much of the initial work had to be redone in favor of something similar to the open-source gpt-2-simple project, maintained by the user minimaxir. With the help of many online guides on how to use GPT-2 and further alter the models provided by it, much of the fine tuning was able to be completed and incorporated into the training class. After all of that was accomplished I began to tweak the training class with smaller and larger sample sizes, altering the size of the data being passed in. Changing the optimizer of the model around to figure out what gave me the best results, as well as using larger and smaller models of GPT-2 with varying learning rates and noise values. With the final iteration having options for altering what kind of sampling to use, 'top k' or 'top p' sampling, to generate an interactive, or unconditional, sample.

2.2 Data

The data gathered for this project was arguably the most intensive and important part. With the restructuring of the direction of the project, much of the initial data mining had to be scrapped in favor of handpicked and formatted data. The data itself is a collection

Table 1: Satisfaction with initial generated sample texts.

Familiarity With Content	Score	Comments
`Not at all`	78	Very neat, some typos though.
`A little`	62	Alright story cohesion, wrong names.
`A lot`	48	Got the flow of the stories right, most everything else was slightly off.

of twenty-three different novels from the publisher The Black Library. This publisher is commonly known as the publisher for the Warhammer 40K novels. I was able to find the novels online in .pdf files and as such ripped all of them and converted them into .txt files for further formatting.

Each book, after being converted into a .txt file, was then formatted to remove all of the unwanted pieces. That is the table of contents, acknowledgements, sneak peak at a future release, the table at the beginning with all the names of people in the book, and what their role is, etc. After all of the unwanted bits were removed and I liked the initial format of the data I ran some initial tests and training through the model. Unfortunately the samples weren't quite what I wanted and as such the formatting of the data continued further.

In the end the data had to be formatted with end of text markers at the beginning and ending of every chapter of each book, as well as pulling the sections of books that are split into multiple parts out into separate documents. Certain words also had to be weighted differently, or have some of them removed from the data, within the model as some of the initial samples that were generated would only use one of the more important character's names for every person that was talking within the sample.

After all of the data had been setup in the way I wanted and found to give me the best results after countless trial and error I tried to see what all it could be used for. In the end about half of the data I gathered was likely garbage and could be tossed out, in the end you could likely get similar results with a much smaller dataset that is much more refined and generalized. As the dataset that I had put together was essentially just one large book written by thirteen different authors, and who knows how many editors, that spanned well over 3,000 standard book pages. However, it is likely that this dataset could be used in conjunction with other book datasets to get more generalized book samples, rather than of this one specific sci-fi genre.

With all of that in mind, another key area I used for data gathering was using a small group of people to read and test the outputs of the models to determine how accurate they were, and how much they enjoyed the samples. This took course over about three weeks, with responses in Table 1 being from the partially trained initial model and dataset before all of the data had been properly formatted and separated out into their respective files.

After all of the data had been formatted and the models had been retrained and tested I went back to the same group of people that I had initially gotten feedback from on the sample generations and asked if they would test it yet again. Across the board the data gathered from the group was that the new model was much more accurate and three of the people asked had stated that they practically couldn't tell that a machine had wrote the small snippet of text they had just read and judged. The respective scores and noteworthy comments can be seen in Table 2.

Table 2: Satisfaction with fine-tuned generated samples.

Familiarity With Content	Score	Comments
`Not at all`	89	Couldn't really tell if it was generated, aside from one or two typos or repetitions.
`A little`	81	If you'd told me this was just from one of the worse novels in the series I would have believed you.
`A lot`	71	Much better than before, still some issues with when names are used and by who.

2.3 Model

My initial question for this entire project was, can I get natural language generation to produce a random sample of text that would be able to entertain your average person as well as trick them into thinking that a person had written the sample of text. With that in mind, the question had to adapt as the initial question was too broad in trying to simply entertain and fool a person into thinking a person had written the sample text. This is largely due to the fact that trying to entertain someone with randomly generated text is incredibly difficult to do. This is due in part to trying to simply get the model to generate somewhat accurate text that doesn't repeat itself over and over again, as well as training against typos within the sample text.

How I ended up having to model the question is, what is important to the GPT-2 model, i.e. what does it actually put weights on and how does it do that? After that was determined, and I figured out how it likes to try and structure sentences, I started examining some of the sample outputs from a completely clean model that had not been trained on any of my datasets. It turns out that GPT-2, at least in some of the smaller models that are available to most users, likes to put a lot more weight on descriptions rather than actual talking between two different people.

This helped to explain why a lot of the samples that were generated during training of the models would very easily mess up when the sample tried to have two or more characters talking in the same scene. This was later remedied by structuring the model slightly differently, by including a section within the data that was nothing more than just conversations between multiple people with no other descriptors, to get it to try and pick up on learning more of how actual conversations flowed within the data.

2.4 Approach

The general approach to the model was to figure out how GPT-2 actual functions under the hood. As this was vital to getting better sample output, as there were some interesting findings when looking within the open-source code libraries of GPT-2. With the interesting bits being what was discussed previously of how GPT-2's smaller models struggle with generating coherent conversations between multiple characters at the same time.

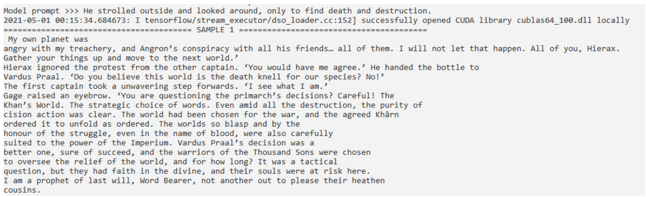
After that had been determined I figured the easiest way to help train the model to handle that exact situation better would be to specifically add to the dataset sections of just conversations between multiple characters. This helped a fair amount as it allowed for samples to keep the names of characters much more consistent when a sample of conversation was generated. However, there were still some issues as it would then start generating repetitions more often as well as some typos in fairly easy words, such as soldier becoming syldor.

With that, I decided the best thing would be to start playing around with weightings on certain words or phrases as well as altering the noise level and the learning rate of the model itself. After tweaking these values for a couple of days I determined that it was time to try and see exactly how well the model would perform when compared with larger models. This was done by training four separate GPT-2 models on the dataset for one week. The models used were the 124M, 355M, 774M, and the 1.5B model. Each of these models performed fairly well, with the 124M and 355M being fairly over trained at the end of the week due to their much more rapid iteration times.

In the end, I had hoped to determine a model that was irrefutably better for my application and found that each model had their own unique benefits as well as their own unique problems. Such as in the 124M model one benefit was the rapid iterations, allowing for quick training and testing. However, it also meant it was easy to overtrain and still had issues with repetition and naming. The 355M had fairly moderate training iteration times, with much better results in coherency and structure than the 124M model. The 774M model was arguably the best model I had used to train, as it gave almost no typos and near perfect coherency for parts of the samples, but was so utterly slow to train that I had to ditch the model in the end. The 1.5B model was very similar to the 774M model, with just about the same issues.

3 EVALUATION

In the end I chose to go with the 355M model as it hit the middle ground on just about all aspects I was looking for to generate my little chapter samples. To help determine accuracy and that the samples being generated weren't just garbage random output I followed the AIDungeon team's validation process. That is, to grab a previously generated sample that you have determined to be excellent, or even just a straight copied section from a random book from a Warhammer 40K novel that wasn't used in the dataset, and then check future generated samples against that. Thus, performing an actual cross-validation check to see just how accurate the generated sample was.



```
Model prompt >>> He strolled outside and looked around, only to find death and destruction.
2021-05-01 00:15:34.684673: I tensorflow/stream_executor/dso_loader.cc:152] successfully opened CUDA library cublas64_100.dll locally
======================================== SAMPLE 1 ========================================
 My own planet was
angry with my treachery, and Angron’s conspiracy with all his friends… all of them. I will not let that happen. All of you, Hierax.
Gather your things up and move to the next world.’
Hierax ignored the protest from the other captain. ‘You would have me agree.’ He handed the bottle to
Vardus Praal. ‘Do you believe this world is the death knell for our species? No!’
The first captain took a unwavering step forwards. ‘I see what I am.’
Gage raised an eyebrow. ‘You are questioning the primarch’s decisions? Careful! The
Khan’s World. The strategic choice of words. Even amid all the destruction, the purity of
cision action was clear. The world had been chosen for the war, and the agreed Khârn
ordered it to unfold as ordered. The worlds so blasp and by the
honour of the struggle, even in the name of blood, were also carefully
suited to the power of the Imperium. Vardus Praal’s decision was a
better one, sure of succeed, and the warriors of the Thousand Sons were chosen
to oversee the relief of the world, and for how long? It was a tactical
question, but they had faith in the divine, and their souls were at risk here.
I am a prophet of last will, Word Bearer, not another out to please their heathen
cousins.
```

Figure 2: An interactive sample generated from user input, using the fine-tuned GPT-2 355M model.

The cross-validation check was done every two-hundred iterations, so every two samples with my setup, and would regularly score around the 0.31 mark for cross-entropy validation loss. To perform this validation a generalized cross-entropy check was done on the sample generated and was then compared to the chunk of information from the dataset that the sample was passed. This helped to generate fairly accurate samples and helped bring the early validation values down from roughly 3.28 to around 0.25-0.35, depending on the model.

My initial question of whether or not a randomly generated text trained on a specific genre of books can provide a convincing sample. Convincing enough to fool someone familiar with the genre and novels into thinking the random sample is a new novel snippet. In the end I believe I answered my initial question as I was able to find some very interesting generated samples, to the point where even I couldn't really tell if the sample hadn't just copied over the chunk of text from the dataset. This is further supported by the fact that I was able to train the models down to a validation loss of just 0.31, which isn't perfect but is far better than I had been expecting to get with my optimization settings. However, I believe that the most supportive part to my model's training is that it was able to fool on average two out of three people from the group of people that I had asked to view and test my project. With most of the people in the group having prior experience and a love for the novel franchise of Warhammer 40K.

With that in mind, some of the limitations of my work is that it was only trained on a small selection of the overall Warhammer 40K novels. Altough I made sure to grade some fairly random novels from the entire collection there were still far too many to use, roughly five-hundred books in total. Then to go through all of those books and clean them up and format them by hand would have been far too strenuous a task for the given time frame of this project. As such, there are some other limitations to my work, mainly that the program itself is setup to work with older versions of GPT-2, rather than the newer GPT-3 model, as well as older versions of tensorflow, v1.12.X-v1.15.0, and not one of the newer v2.X versions of tensorflow.

As for what one can reasonably expect this project, and these models, to produce is a fairly coherent and mostly accurate sample of about one or two pages of a book from the Warhammer 40K franchise. However, they shouldn't expect complete accuracy when comparing the samples generated to what is accurate within the franchise, as there will sometimes be snippets where someone's name is replaced with another's name and the sample doesn't make sense in that context. Such as a sample where, "The Emperor" is talking to, "Magnus" as if Magnus were the Emperor and Magnus is talking to the Emperor as if they were Magnus. This is a fairly common issue within the samples generated on the smaller models. It is still present in the larger trained models but is not present nearly as much as in the smaller models. Therefore, the user should also keep in mind that these models will not really generate accurate and coherent text outside of generating text for Warhammer 40K. It can generate fairly accurate sci-fi samples however that's about it, and even that is stretching it a little bit far as it still wants to stick a lot of the trained terminology from the novels into the samples.

4 IMPLICATIONS

Moving on to who exactly would use my work. It's likely that not that many people would end up using my work as it is very niche and specialized on a specific dataset and most of the tweaks I have made to the model and data are specific to get the kind of output I desired. However, it's possible that some people might use my work, with those people being in the form of other Warhammer 40K nerds that just truly love the franchise and want to try and generate their own little bits of stories. Whether that is for a potential home brew campaign of dungeons and dragons or whether they wanted to generate some kind of sci-fi snippet of text to use in some kind of story.

As for who this project and research might affect, I would likely say only someone who wants to try and work with GPT-2, and even that is unlikely with the continued releases of newer and better versions of the GPT models. For example, it's unlikely that anyone would want to re-use this project as the newer GPT-3 provides much better support, although it comes at the cost of being incredibly bulky and almost impossible to train and modify upon on a single home machine. As such, I believe that someone might be able to get use from this project but it's unlikely unless it is done sometime within the next year or so.

With all of that in mind, it's unlikely this project will affect anyone in a negative way. This is due to the fact that I'm not trying to make fake blog posts with the samples, and potentially infringe on The Black Library's copyright. As well as that I'm not really posting any of the samples anywhere other than using them for neat little personal projects as well as using them for further training of the model until I can eventually get it to a point where no one can really determine whether or not the generated sample is real or fake.

5 CONCLUSION

In the end, I set out on a long project path of trying to build an interactive AI generated choose your own adventure game. As I am only one person, that project was shelved to be worked on at a later date as a personal project in my spare time. The project pivoted to trying to generate natural language that is incredibly difficult to tell apart from the original source, and I believe that I have properly accomplished that goal. Although it might not be perfect, I am able to generate random samples of text that can very closely mirror snippets of text from within a Warhammer 40K novel. To the point that it is almost able to fool people who have read large portions of the Warhammer 40K franchise. As such, it can be incredibly dangerous for people in the future, as there is already software capable of fooling someone into thinking that what they're reading is entirely man made. So, the future can look bleak but it will likely still be the same as instead of some person writing a news article you don't care about, it'll just be a machine writing a news article, or even book, about something that you don't care about.

REFERENCES

[1] Latitude-Archives AIDugenon *Latitude-Archives* https://github.com/Latitude-Archives/AIDungeon

[2] What Marketers Need to Know About AI Content Generation Today *Aritifical Intelligence Institute* https://www.marketingaiinstitute.com/blog/what-marketers-need-to-know-about-ai-content-generation-today

[3] Beginner's Guide to Retrain GPT-2 (117M) to Generate Custom Text Content *AI Labs* https://medium.com/ai-innovation/beginners-guide-to-retrain-gpt-2-117m-to-generate-custom-text-content-8bb5363d8b7f

[4] Minimaxir GPT-2-Simple *Minimaxir* https://github.com/minimaxir/gpt-2-simple

[5] How to Use OpenAI's GPT-2 to Create an AI Writer *Nandeshwar* https://nandeshwar.info/data-science-2/use-openai-gpt-2-to-create-ai-writer/

[6] How to sample from language models *Towards Data Science* https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277

[7] OpenAI GPT-2 *OpenAI* https://github.com/openai/gpt-2

[8] How To Make Custom AI-Generated Text With GPT-2 *Max Woolf's Blog* https://minimaxir.com/2019/09/howto-gpt2/