Magic: The Generating – Building an MTG Language Model

Parker Griep
Northeastern University
griep.p@northeastern.edu

Abstract

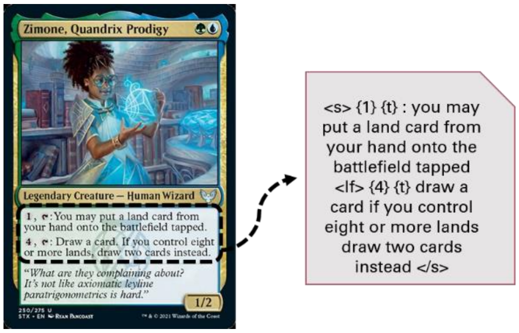
Language models are ubiquitous in today's natural language processing applications. However, recent architectures have increased the number of parameters and the amount of training data required to see effective results [6, 9]. We propose an exercise in building language models on the popular trading card game Magic: The Gathering's rules text. This structured language is small (<100MB) and provides a good resource for students to build more complicated language models with success. Because cards are game pieces and have financial value, their generation is interesting and can engage students.

1. Introduction

Magic: The Gathering is a popular trading card game from 1993 with over 35 million players in 2018 [2]. Its main game pieces are cards that take many forms, with different colors, templates, number of sides, and represent creatures and spells from across the multiverse. Some of these cards have financial value, with one card, pictured below, selling at up to $551,000 in January 2021 [2]. Each card contains rules-text describing what that card can do. Rules-text is structured but varied enough for the game makers to design new cards multiple times a year. For these reasons, Magic: The Gathering rules-text seems well positioned to be the dataset of a language model.

2. Dataset

Our primary dataset for this project is a comprehensive list of all printed Magic: The Gathering cards. We procured these through MTGJSON, an open-source project that stores cards in a portable format [4]. We filtered the dataset down into cards that had text, as some cards only have art, and filtered to cards only legal in the game format Vintage. This helps us avoid problematic cards such as those that depict racist concepts [3]. It should be noted that the amount of data is quite small, with the dataset totaling in at roughly 21,000 samples and a training set size of 16,000.


Zimone, Quandrix Prodigy
Legendary Creature — Human Wizard
<s> {1} {t} : you may put a land card from your hand onto the battlefield tapped <lf> {4} {t} draw a card if you control eight or more lands draw two cards instead </s>

2.1. Preprocessing

Since our work only observes rules-text, we ignore other attributes of a card. These include Card Name, Mana Value/Cost, Power/Toughness/Loyalty, etc. To prepare the rules-text for training, we use the following steps. Given an entire card's rules-text, we first normalize the text. We accomplish this by lowercasing the text and substituting self-references in a card. Next, we substitute special punctuation and line feed characters with special tokens. Lastly, we space out elements in mana costs and ability activations (e.g. "{1}{t}:" => "{1} {t} :").

3. Methods

3.1. Model Choices

For this project, we compare two recently state-of-the-art model architectures: GPT2 and a multi-layered Bidirectional LSTM. Our priorities are to choose models that have succeeded in language modeling, models that are easy to configure, and quick to setup. We use HuggingFace's implementation of GPT2 and used the Keras library to implement the LSTM network.

3.2. Tokenization

To simplify comparisons between models, we use one tokenizer to standardize the vocabulary of the dataset. We use a ByteLevelBPETokenizer implemented by HuggingFace's tokenizers library. The tokenizer utilizes a technique called "Byte-Pair Encoding" which breaks up rarer, larger words into subword units [8]. Our final vocabulary size after training the tokenizer on all cards is 2,936 types with 27 special tokens. These special tokens include a beginning of card token, end of card token, a padding token, a mask token, and an unknown token.

Model	**Tokenizer**
vocab_size	2,936
num_special_tokens	27

3.3. GPT2

OpenAI's GPT2 model was trained as a language model for the Internet. It is the successor to OpenAI's previous model GPT. It relies on a multi-layered transformer architecture with self-attention layers [6]. We use this model since it performs well with complex language on the Internet, and we want to see how it adapts to a more structured, limited language. To compare to our other model more easily, we train GPT2 from scratch, rather than fine-tuning an existing set of weights.

We use HuggingFace's implementation of GPT2 because it is relatively straightforward to setup. We also use a DataCollator to automatically build sequences from samples, padding when necessary. We also enable masking on some percentage of the tokens, such that multiple epochs are less likely to overfit. The architecture for our GPT2 model can been seen below. Due to time constraints, we did not tune hyperparameters.

Model	**GPT2**
num_positions	512
num_context	512
num_embeddings	128
num_layers	8
num_attention_heads	8

3.4. LSTM

Before recent innovations with transformer-architectures, LSTMs were the leading language modeling network. Popular architectures that rely on bidirectional LSTMs include ELMo [5]. While ELMo uses contextualized embeddings, our model trains on non-contextualized word embeddings. LSTM or Long Short-Term Memory networks are like RNNs except they have a linear cell-state that controls memory instead of a non-linear state. This linearity helps avoid the vanishing/exploding gradients problem [5].

We use the popular library Keras, which is built on top of TensorFlow, to implement this model. The model architecture can be found in the figure below. Like GPT2, we did not tune hyperparameters.

LSTM Layers
Embedding
BidirectionalLSTM
BidirectionalLSTM
Dropout
Dense
SoftMax

Model	**LSTM**
num_embeddings	248
num_lstm_units	30
num_layers	2
dropout	0.3
sequence_length	10

3.5. Evaluation

To evaluate our models, we use the popular metrics BLEU and ROUGE [1, 7]. BLEU (Bilingual Evaluation Understudy) measures the quality of text from a translation. The score is bounded between 0 and 1, where values near 1 correspond with realistic text. The other metric, ROUGE (Recall-Oriented Understudy for Gisting Evaluation), counts the number of overlaps for a number of N-grams. For example, ROUGE-1 references to unigrams. To acquire examples to feed into these metrics, a hold-out set was taken from the dataset. We use the first (30%-60%) of each validation sample as a prompt and compare the model's output to the real card.

Additionally, we also built a survey to capture more qualitative results. We generated 4 unprompted random cards from each model, as well as randomly chose 4 real MTG cards, and turned them into survey questions. Each question was, "How realistic is this generated MTG card?" Each respondent was debriefed that there were 4 real cards after they had completed the survey. Respondents answered for each card by choosing a number on a 1-7 Likert scale, where 1 represented "Unrealistic" and 7 represented "Realistic". A link to the form can be found at `https://forms.gle/A2VKdG8JfPtwtb92A.`

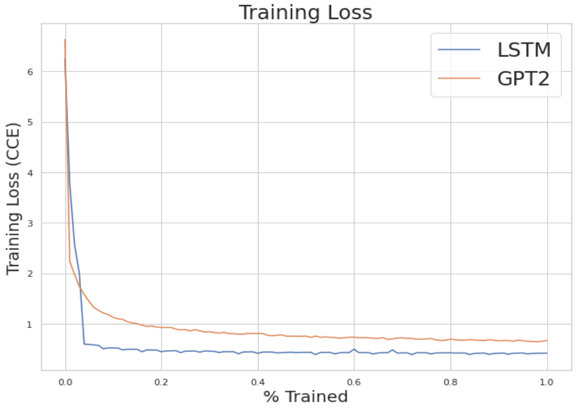
4. Results

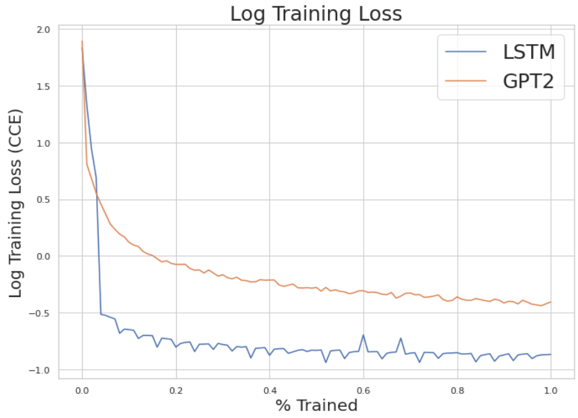
Model	# Parameters	# Epochs	Total Train Time
GPT2	2,027,776	500	5h12m42s
LSTM	996,024	25	1h15m25s

4.1. Training Comparisons

Both models were trained on consumer hardware: Ubuntu 20.04, i7-9700k 3.6Ghz, GTX 1070. However, the GPT2 model took nearly 5x as long to train as the LSTM. Inspecting the tables for training loss and log training loss, we notice that the LSTM reaches convergence faster than GPT2 and with a lower loss. However, training loss is not the best metric to analyze since a non-negligible amount of the predictions are on pad tokens, inflating the accuracy / deflating the loss.

As a sanity check, many of the generated cards were cross-checked with real cards to ensure the model was not overfitting. Less than 5% of generated cards from both models have the same text as real cards, however, in most of these cases the card's text is quite short (~3-5 tokens).





4.2. BLEU and ROUGE

Better metrics for model analysis include BLEU and ROUGE. From our results, we can conclude that the LSTM model should correspond with more realistic MTG cards. Oddly, the LSTM outperforms GPT2 by almost every metric except the ROUGE-L Recall, which means that GPT2 is producing more relevant subsequences of cards.

	BLEU		
Model	**Score (μ)**	**Precision (μ)**	**Length Ratio (μ)**
GPT2	0.227	0.232	2.233
LSTM	**0.354**	**0.363**	**1.355**

	ROUGE-1		
Model	**Precision (μ)**	**Recall (μ)**	**F1 (μ)**
GPT2	0.321	0.687	0.409
LSTM	**0.546**	**0.621**	**0.535**

	ROUGE-2		
Model	**Precision (μ)**	**Recall (μ)**	**F1 (μ)**
GPT2	0.237	0.479	0.298
LSTM	**0.432**	**0.454**	**0.411**

	ROUGE-L		
Model	**Precision (μ)**	**Recall (μ)**	**F1 (μ)**
GPT2	0.309	**0.663**	0.394
LSTM	**0.538**	0.610	**0.527**

4.3. Survey Results

For survey results, our target was to have a model achieve a mean above the middle Likert value (in this case 4). We did achieve this goal with one model. Out of 12 respondents, they found cards produced by the LSTM more realistic than unrealistic with an average score of 4.23. The GPT2 model has a mean score of 3.807, relatively close to our goal. However, the standard deviation of both models sits at 2 points, meaning that the models could produce wholly realistic cards and completely unrealistic cards.

An interesting note about some of the most likely outputs. The LSTM model seems to be exhibiting knowledge of the relationship between color and effect. The mana color red ({r}) often deals with damage, which was generated here.

5. Conclusions

We were able to successfully build 2 language models for Magic: The Gathering rules-text that produce novel, realistic cards. These models can both be run on consumer hardware and operate on a small dataset. We assessed the quality of these models through multiple metrics and a qualitative survey. We conclude that Magic: The Gathering cards are an excellent domain for training recently state-of-the-art language models.

6. Future Work

For future work, we would begin tuning the hyperparameters of each model to see if we could produce more realistic output. More interestingly, we might also start to add more information besides the rules-text into generation. In theory, we could encode all a card's properties as text and feed it to these models to generate wholly new cards.

	Likert Score [1, 7] **Q:** How realistic is this generated MTG card?		
Model	**(μ)**	**(σ)**	**Highest Scoring Sample**
Ground Truth	**5.827**	**1.34**	<s> whenever you gain life put that many +1/+1 counters on CARDNAME </s>
GPT2	3.807	2.094	<s> choose one <em> <lf> - all creatures get -2/-2 until end of turn <lf> - CARDNAME deals 1 damage to each creature <lf> - CARDNAME deals 2 damage to target creature </s>
LSTM	4.23	2.366	<s> {r} sacrifice a creature : CARDNAME deals 2 damage to any target </s>

7. Acknowledgements

We would like to thank Felix Muzny for their excellent teaching in Northeastern University's course CS4120: Natural Language Processing. Additionally, we would like to thank TAs Latika Korad and Joshua Alter for making our virtual semester run smoothly.

8. Supplementary Material

The code for our project can be found at `https://github.com/Buntry/cs4120-nlp-final`

References

1. "BLEU." In *Wikipedia*, November 9, 2020. https://en.wikipedia.org/w/index.php?title=BLEU&oldid=987822284.
2. Bolding, Jonathan. "With a Black Lotus Sold at $500k, Magic: The Gathering Hits a New Level." *PC Gamer* (blog), January 30, 2021. https://www.pcgamer.com/with-a-black-lotus-sold-at-dollar500k-magic-the-gathering-hits-a-new-level/.
3. MAGIC: THE GATHERING. "Depictions of Racism in Magic." Accessed April 18, 2021. https://magic.wizards.com/en/articles/archive/news/depictions-racism-magic-2020-06-10.
4. "MTGJSON.Com | Cataloging All Magic: The Gathering Cards in Portable Formats." Accessed April 18, 2021. https://mtgjson.com/.
5. Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations." *ArXiv:1802.05365 [Cs]*, March 22, 2018. http://arxiv.org/abs/1802.05365.
6. Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models Are Unsupervised Multitask Learners," n.d., 24.
7. "ROUGE (Metric)." In *Wikipedia*, September 3, 2019. https://en.wikipedia.org/w/index.php?title=ROUGE_(metric)&oldid=913825951.
8. Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units." *ArXiv:1508.07909 [Cs]*, June 10, 2016. http://arxiv.org/abs/1508.07909.
9. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." *ArXiv:1706.03762 [Cs]*, December 5, 2017. http://arxiv.org/abs/1706.03762.