Grover: a new voice

Matthew McAvoy

[2020-05-11 Mon]

Contents

1 Introduction

The goal of this project is to investigate the ability of a pre-trained model to fine tune on new data to speak in an author's voice and quantify a lower bound on this task.

This project was inspired by the publication of GROVER [1], where researchers extended GPT-2 [2] by adding trainable fields that include not only the body of a text, but also the domain, date, author, and headline of an article; with the hope that conditional text generation will be more realistic than unconditional. What they found is that Grover is able to generate articles as if they were written by humans.

One aspect that is not convered in the Grover paper is how well the model trains on a conditioned variable, for example the author of the article. The implications of this are significant. If Grover or a similar algorithm can be trained to speak in the voice of an author and have it say something contrary to their moral values could be quite damaging. To better understand how Grover might be trained to perform such a feat is the aim of this project. While Grover is ideal, due to limitations in the ability to fine-tune the model, GPT-2 was used as the algorithm for fine-tuning. This loses out on specific layers included during training yet as shown in the results section, some of the conditionality can be recovered through proper format of the training dataset.

With specifics covered later, a human-interpreted check on the algorithm appears to be within 300 examples. That is, with only 300 samples of text on a specific author, can GPT-2 speak in that author's voice.

2 Methods

2.1 Twitter data collection

Data was collected using the Twitter api. Initially a thirty day window of up to 100 tweets per user per day was collected. The users can be seen in table 1. An additional 35 days were collected for nytimes. Once tweets were collected, they were formatted by adding tags around pertinent fields. A separate file for each user was also saved where the text was cleaned by means of: substituting all urls with a url tag, all @user with a user tag, all #hashtags with a hashtag tag, removing non-ascii characters, and lowercasing. The tokenizer then had these tags added to it.

```
special_tokens_dict = { ’eos_token’: ’<|endoftext|>’,
’additional_special_tokens’: [
’<|begindomain|>’, ’<|endofdomain|>’, ’<|begindate|>’,
’<|endofdate|>’, ’<|beginauthors|>’, ’<|endofauthors|>’,
’<|begintitle|>’, ’<|endoftitle|>’, ’<|beginarticle|>’,
’<|endofarticle|>’, ’<|url|>’, ’<|user|>’, ’<|hashtag|>’
] }
```

Due to a complication that wasn't resolved until later, only nytimes had additional data collected and a comparison of the effect of cleaning the tweets were performed.

2.2 Congress data collection

A member of the Lunar Lab at Brown University provided a reference for an additional data source, a dataset containing text from the United States Congressional Record from the 43rd to 111th Congresses [3]. Two sessions, the 61 and 111 congress were chosen and the top two republican and democratic congressman were selected and filtered to create a subset of data for easier analysis. The texts were similarly cleaned as tweets with tags included indicating the person speaking. An additional step of removing procedural words were performed. The words removed can be seen below by use of regex:

```
procedural_words = [
    'yield', 'motion', 'order', 'ordered', 'quorum',
    'roll', 'unanimous', 'mr\.?', 'mrs\.?', 'speaker\.?',
    'chairman\.?', 'president\.?', 'senator\.?', 'gentleman\.?',
    'madam\.?', 'colleague\.?', 'today', 'rise',
    'pleased to introduce', 'introducing today', 'would like',
    'i suggest the absence of a']
```

2.3 formatting data for conditionality

A tweet to mimic Grover's conditional fields have tags surrounding parameters accumulated/generated from a tweet. The fields included are: domain, date, author, title, and article. For cleaned tweets, url, hashtag, and users are also included as place holders to standardize a tweet. Title is the only generated field as a concatenation of the author and the date. One such tweet would look like below, note that an entire tweet is normally on one line to facilitate training, spaces have been included for readability:

> <|begindomain|> twitter <|endofdomain|>
> <|begindate|> 03-30-2020 <|endofdate|>
> <|beginauthors|> The New York Times <|endofauthors|>
> <|begintitle|> The New York Times Mon Mar 30 03:30:06 +0000 2020 <|endoftitle|>
> <|beginarticle|> two of the u.s.s largest health insurers agreed to protect their customers from out-of-pocket costs if they need t <|url|> <|endofarticle|>
> <|endoftext|>

Similarly, a congress text can be seen below. Note that there are not as many fields possible to include:

```
<|beginauthor|> 111120961 <|endofauthor|>
<|beginspeech|> i ask consent that the text of the joint resolution
be printed in the record. <|endofspeech|>
<|endoftext|>
```

2.4 training the model and generating text

Huggingface [4] provided an easy to use interface to GPT-2. There are various sizes of GPT-2 available, the one used was the small model with 117 million parameters. The model was downloaded and fine-tuned using the processed data. Checkpoints were taken every 100 steps for nytimes and every 200 steps for congress. A batch size of two was used for nytimes and one for congress. The default value of four would have been preferable, however due to limitations on GPU ram, the values chosen were the largest possible at the time of experiment. A separate tokenizer for nytimes and congress including the fields were used.

After training, the models generated text. For each checkpoint that had generation, three different seed values were used; the seed value is essentially a randomization parameter that when it's the same with the same prompt, will produce the same text, and when it's different yet with the same prompt, will produce different text. The prompts used were the author field surrounded by each of the selected authors.

3 Results

With no defined way to quantify how well an AI generated text sounds like someone, the next best alternative is to human-verify it by reading and comparing. One empirical step that was used is a check to see how well the model generates the tags to surround various fields. Stabilization is defined here as when the algorithm will produce tags surrounding fields in a consistent manner from some checkpoint onwards. Through these methods we can now inspect some generated text.

3.1 nytimes

Checkpoint-300 uncleansed

```
<|beginauthors|> nytimes <|endofauthors|>
The New York Times <|endofauthors|>
<|begintitle|> The New York Times Wed Jan 23 17:30:05 +0000
```

> 2020 <|endoftitle|>
> <|beginarticle|> Trump and his allies are making continuing campaign rhetoric against the powers of the presidency. They're living up to their motto of https://t.co/DkPXKa9WOm7

Checkpoint-300 cleansed

> <|beginauthors|> nytimes <|endofauthors|>
> The New York Times <|endofauthors|>
> <|beginarticle|> The New York Times Thu Mar 23 09:50:12 +0000 2020 <|endoftitle|>
> <|beginarticle|> polls are at an eye-popping pace today with momentum around the world from japan to canton qui <|endoftitle|> <|url|>

Checkpoint-3000 uncleansed

> <|beginauthors|> nytimes <|endofauthors|>
> 02-04-2020 <|endofdate|>
> <|beginauthors|> The New York Times <|endofauthors|>
> <|begintitle|> The New York Times Tue Feb 04 21:30:03 +0000 2020 \<|endoftitle|>
> <|beginarticle|> A spokesman for Mr. Trump indicated he would abandon plans to travel to Mexico if the U.S. determines https://t.co/Nqo6OtTfhF <|endofarticle|>

Checkpoint-3000 cleansed

> <|beginauthors|> nytimes <|endofauthors|>
> 02-26-2020 <|endofdate|>
> <|beginauthors|> The New York Times <|endofauthors|>
> <|begintitle|> The New York Times Sun Feb 26 23:00:01 +0000 2020 <|endoftitle|>
> <|beginarticle|> the peoples fault in my case is clear. we delivered a drone weapon with a deadly and unpredictable resonance. today, they want to push <|endofarticle|>

At checkpoint-300, legible text appears to begins to form. At checkpoint-700, uncleansed text begins to converge. At checkpoint-1300 cleansed text begins to converge. After convergence, some standadard statistics were calculated on the text, seen in the table 1

Table 1: nytimes convergence rate and sample statistics

cleansed	convergence checkpoint	min	mean	max	std deviation
No	700	5	21.86	35	6.3
Yes	1300	14	22.08	33	3.98

3.2 congress

While four speakers were trained during each session, only one is displayed here for brevity. The difference between the generated text for each of the speakers appeared to be negligible..

3.2.1 Session 061

Checkpoint-400

> <|beginauthors|> 61076130 <|endofauthors|>
> that does not imply that there would be time for the committee to go on discussing abolition. for which the audience would have been very thankful to the from samuela. <|endofspeech|>

Checkpoint-2000

> <|beginauthors|> 61076130 <|endofauthors|>
> I demand the following question: in it is made my question. where the reading has been carried in the amendments to the amendments to which i consent. i have stated something of what has been done to pay the debt accumulated after in line 9. and i stated in this bill that one of the payment for salaries and benefits. which is borne out of the annual appropriation provided for by the proviso. or the debt can be collected without any provision given by law. but after the appropriation is deposited in the account. from which money is collected as required and obtained by the appropriation. those expenses. but as owing, are eliminated. i objection to them at all. <|endofspeech|>

Checkpoint-4000

> <|beginauthors|> 61076130 <|endofauthors|>
> 15.100.000 </. are there any parties in the district of columbia who had any difficulty in educating the chicago street people? <|endofspeech|>

Checkpoint-6000

> <|beginauthors|> 61076130 <|endofauthors|>
> i at last have a revised. or improved. copy of the bill. <|endofspeech|>

3.2.2 Session 111

Checkpoint-400

> <|beginauthors|> 111120391 <|endofauthors|>
> hi there. i will be taking the liberty to ask there to be no further reading. <|endofspeech|>

Checkpoint-2000

> <|beginauthors|> 111120391 <|endofauthors|>
> now. if youre curious. youre read all the evidence on the ground and the evidence on the ground is overwhelming. to my knowledge. the agreement has never been negotiated. it is clear that i will try to prevail on this issue. i am not going to stay here. though. so we have a long sallie rule that basically prohibits sidebarring monday. the consent decision will be final. the majority will appoint the next justice. and we can work this out. i think we are a couple of minutes done on that one. it is very important for us to have that one together. <|endofspeech|>

Checkpoint-4000

> <|beginauthors|> 111120391 <|endofauthors|>
> i have been with the republican for 7 years. it is my personal opinion that the statement that was sent is misleading. the author has been in contact with several republican senators and has been informed of my statement. for instance. i just sent an email to the republican leader at the beginning of this month. the of the armed services committee wrote it on june 4. he is asking for its first reading. (...) it would take 5 hours to actually vote on the amendment. i am giving my own opinion. this was written by my friend from illinois. he indicated that. following this soothing statement and the amendment. that. in

Checkpoint-6000

> <|beginauthors|> 111120391 <|endofauthors|>
> my only objection to the concurrent resolution is that it does not give us a chance to get to the bill by time this week. we just had a conference call yesterday with representatives from all of these states who have served on committees and committees in s meeting to see if there are anything that they think would be helpful. and the process is getting pretty tedious as they run it. but we are coming up with a bill that is both administratively and fiscalarily important. it is an effort to do things that work. a legislative effort to deal with these problems. and the bill will provide me with a chance to present this caucus or meet with someone who has served on it and reach an agreement on what the bill would be. <|endofspeech|>

We don't see as well convergence, that is formatted tags included in the generation. As well, even after 10,000 examples, there were still generated text that were short or non-sensical. The 061 session had a higher occurence of less syntactically formed sentences, one hypothesis for this is it's a byproduct of the language of the age and the algorithm not being trained with any text from that period. An additional observation can be seen in session 111 checkpoint-4000. This author is John McCain, who was a prominent republican senator. We see the generated text includes words 'republican' and 'senator', indicating it's picking up on his voice and attuning to it.

4 Conclusion

We see that the algorithm appears to be able to produce readable text with as few as 300-400 examples. The content and format of the data affects the convergence rate. Comparing nytimes to congress, consistent and structured fields allows an easier ability for the algorithm to generate structured text. The effect of cleansing the text slows this convergence yet produces potentially a more generalized form with generic tags. Finally, the algorithm is able to speak in different voices, represented by the comparison of the congress sessions 061 and 111.

Future work will look to accumulating more specific data on a single topic where the authors will each have distinct voices with the goal of making it easier to compare the voice of the generated text. An additional avenue of research is how the neurons in the algorithm affect fine-tuning, are all or nearly all neurons pertrubed during fine-tuning, or are only a few with significant value changes?

We conclude by re-iterating the potential implications of these algorithms. The original authors of both GPT-2 and Grover pursued their research to better understand the implications developing these language models could have, both for positive and negative impact. The prospect of someone able to imitate someone else in a convincing way, with such a low amount of data needed for fine- tuning is ripe for malicious use. It is the hope of this avenue of research to better understand how well these algorithms can fine-tune and at a later time, develop a quantitative method to identify generated false text from original text from an author.

References

[1] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending Against Neural Fake News, 2019.

[2] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[3] Jesse M. Shapiro Gentzkow, Matthew and Matt Taddy. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. *Stanford Libraries*, 0, 01 2018.

[4] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.