Decoding Strategies for Interactive Narrative Generation

Alexandra DeLucia
aadelucia@jhu.edu

Aaron Mueller
amueller@jhu.edu

Lisa Li
xli150@jhu.edu

Abstract

Narrative generation is an open-ended NLP task which employs similar methods to neural response generation. However, despite the prevalence of response generation research, narrative generation has received much less attention and has generally not employed innovations from this related subfield. We aim to bridge this gap by applying and evaluating recent advances in decoding methods for neural response generation to neural narrative generation. In particular, we perform ablations across model size, nucleus sampling thresholds, narrative lengths, and diverse decoding hyperparameters. We find that (1) GPT-2 outperforms previous seq2seq models made for this task, and that larger versions are better; (2) nucleus sampling is generally better than top-k sampling; (3) there is a positive correlation between p and narrative interestingness, but a negative correlation between p and story coherence; and (4) diverse decoding generally makes stories more interesting, but at the cost of reduced fluency.

1 Introduction

Narrative generation (or interactive story generation) is the task of generating a creative output response given a user-defined input prompt. This output can be a sentential story closure, a paragraph, or a structured story with multiple paragraphs. This input and output setup is similar to the dialogue generation task of chatbots, for both tasks convert some variable-length sequential input from a user to an automatically generated variable-length sequential output. Thus, the neural models and methods proposed thus far for story generation and dialogue generation have been similar.

However, as story generation is largely focused on coherence across long outputs, the strategies used in this subfield have evolved separately from those in chatbot response generation; the latter has been more concerned with generating interesting and diverse—and typically *short*—outputs. This means that, while many beneficial techniques may have arisen from one niche, they are not often employed in the other. Thus, this paper aims to apply techniques from neural response generation to narrative generation in order to investigate the potential benefits—and pitfalls—of applying newly developed decoding techniques in this new domain.

To this end, we investigate the following phenomena:

1. The correlation between model size and output quality using the current generative state-of-the-art, GPT-2.
2. The relative performance of top-k and nucleus sampling in the narrative generation domain.
3. The effect of the nucleus sampling threshold p on narrative quality.
4. The effect of diverse decoding techniques (namely, the maximum mutual information objective) with various diversity strengths λ on narrative quality.

2 Related Work

There have been multiple recent efforts to build narrative datasets. Two very large collections of online books are Project Gutenberg[1] and Smashwords.[2] Project Gutenberg is a collection of classics and other out-of-copyright stories freely available to the public, and Smashwords contains stories from current self-published authors. Many researchers have also made efforts to gather story datasets, as well as designing generative models for stories. Fan et al. (2018) train a model that first generates a prompt and then transforms it into a passage of text;

[1] https://www.gutenberg.org
[2] https://www.smashwords.com

they also collect the WRITINGPROMPTS and ROCSTORIES datasets. Mostafazadeh et al. (2016) is a collection of five-sentence stories on which the authors propose a method of Story Cloze Evaluation, which checks a pre-trained model's ability to predict what should happen next in a story given prior context. The bAbI Project (Weston et al., 2015) contains many tasks, the most relevant of which is a task on children's books (Hill et al., 2015) where the evaluation entails casting a Cloze story completion to a multiple choice problem. Zhu et al. (2015) introduced the BOOKCORPUS dataset, and their model generates paragraph captions for movie scenes.

Inference is another important component in narrative generation. In general, there are two streams of work on promoting quality and diversity of decoding methods. One direction is to explore various decoding objectives, including the standard log-likelihood loss and some novel objectives: Li et al. (2016a) introduce an objective that maximizes mutual information (MMI) between the source and target. They propose a bidirectional (MMI-Bidi) and an anti-language model (MMI-antiLM) objective; in this work we evaluate the effect of the MMI-antiLM objective on response generation diversity and quality. Nakamura et al. (2018) use Inverse Token Frequency to reweight generated tokens. Xu et al. (2018) and Zhang et al. (2018) use adversarial loss to optimize for diversity, informativeness, and fluency.

Another direction of research focuses on the algorithm used for search during decoding: beam search has been a popular inference algorithm to approximate the argmax of some decoding objective that is factored by token. At each time step, the beam search algorithm keeps track of the top B beams. When $B = 1$, this method reduces to the *greedy* decoder, which chooses the argmax over the model's token distribution at each time step. Some extensions to beam search include "noisy parallel approximate decoding" (Cho, 2016) which perturbs the hidden state; "top-g capping" (Li et al., 2016a,b) which encourages choosing from diverse candidates; "iterative beam search" (Kulikov et al., 2019) where beam search is run multiple times and previous intermediate states guide the search space of new iterations; and "clustering post-decoding" (Ippolito et al., 2019) which assigns candidates to clusters, ideally preferring novel clusters.

An alternative to search is **random sampling**, which samples from the (typically constrained) model distribution at every time step. Such methods include "top-k" which restricts the sampling space to the top k most probable tokens at every time step (Fan et al., 2018; Radford et al., 2019) and "nucleus sampling" which drops the long tail by thresholding the distribution (Holtzman et al., 2020). We evaluate both of these methods but focus primarily on nucleus sampling.

3 Experimental Setup

3.1 Dataset

For our task of interactive narrative generation, we follow the lead of Fan et al. (2018) and train on their long-form response dataset WRITINGPROMPTS.[3] This dataset was built from the subreddit `r/WritingPrompts`[4], where users post a "prompt" consisting of a few sentences, and then other users reply to the post with a story continuing the prompt. An example prompt follows[5]:

> A group of young women, all from wealthy aristocratic families, are sent off to attend a prestigious 'Finishing School' to become proper ladies of society... Only to find themselves at a prestigious military academy instead.

To create datasets of varying lengths—and to make the dataset compatible with our model (GPT-2, discussed more in Section 3.2)—we preprocess the WRITINGPROMPTS dataset as follows:

1. Remove all prompts that are not tagged with `[ WP ]`. Other tags in `r/WritingPrompts` have different meanings and response requirements, such as having to occur in an established universe; we want only unconstrained responses. This is because it would be unfair to judge the model's performance in generating a story in, say, the Harry Potter universe, if it does not have the necessary background of all the Harry Potter books. The same goes for the other prompts that specify constraints, such as

[3] https://github.com/pytorch/fairseq/blob/master/examples/stories/README.md

[4] https://www.reddit.com/r/WritingPrompts/

[5] https://www.reddit.com/r/WritingPrompts/comments/fr4d4n/wp_a_group_of_young_women_all_from_wealthy/?utm_source=share&utm_medium=web2x

not being to include specific words. Further, we want the responses to be as creative as possible, so no writing restrictions are ideal.

2. Replace the `[ WP ]` tag at the beginning of the prompts with `[WP]` (i.e., make it a single-token tag) and add a `[RESPONSE]` tag to the beginning of the responses. We define these as special tokens in GPT-2's tokenizer so that these tags are not split into subword units.

3. Create small, medium, and long versions of each response by using all content from (1) before the first line break, (2) before the third line break, and (3) the entire response, respectively. It is important to note that only the response length is changed and they all have the same number of prompts. These are herein referred to as the "small", "medium", and "large" datasets/response lengths, and are treated as separate corpora. Thus, we have 3 train, validation, and test corpora total.

4. Combine the source (prompt) and target (response) strings into one string and write to a text file, where each prompt-response pair is delineated by a newline character.

5. Add `<|startoftext|>` and `<|endoftext|>` tokens to the beginning and end of each prompt-response pair, respectively.

During step 3, we create multiple versions of the training set with different length responses to evaluate the quality of narrative generation for outputs of various lengths. During pre-processing, we had originally intended to use the first sentence, first three sentences, and entire response for the small, medium, and large datasets, respectively. However, we found that these were not good approximations of story length, since sentence segmentation is often inaccurate and the length of the first sentence is often too short to be interesting. Thus, we use line breaks instead of sentence boundaries to create these different datasets. In a random sample of 100 responses from the training set, we found that the median number of sentences between line breaks is 2 to 3; thus, these correspond roughly to half-paragraphs. See Table 1 for the sizes of these datasets.

3.2 Narrative Generation with GPT-2

Instead of the neural-based models used in Fan et al. (2018), we focus on the generative Transformer-based model GPT-2 (Radford et al., 2019).[6] We employ this model because it is current state-of-the-art in text generation.

GPT-2 is a Transformer-based model that is a modification of OpenAI GPT (Radford, 2018). The changes include sub-block layer normalization, as well as a larger vocabulary (50,257), context size (1024), and batch size (512). The model comes in different sizes: *small* (117M params, 12 layers, 768 dim), *medium* (345M params, 24 layers, 1024 dim), *large* (762M params, 36 layers, 1280 dim), and *extra-large* (1542M params, 48 layers, 1600 dim). We use the small and medium GPT-2 models for output quality comparison. The large model, unfortunately, was too large and was not feasible to train on the medium and large datasets, even on a Google Cloud instance with multiple Tesla P100 GPUs.[7]

GPT-2 is originally trained on WebText. For this work, we fine-tune both GPT-2 models (small and medium) on the small, medium, and large version of the WRITINGPROMPTS dataset discussed in Section 3.1. Each example in each corpus is of the format "`<|startoftext|>` `[WP]` PROMPT `[RESPONSE]` RESPONSE `<|endoftext|>`". Fine-tuning is performed on Google Cloud instances with a NVIDIA Tesla K80s and T4s.

3.3 Decoding Methods

After GPT-2 is fine-tuned on the WRITINGPROMPTS dataset, we evaluate the model's generated responses with a parameter sweep of p for nucleus sampling. We also provide a small comparison with top-k sampling in Section 4.1. As previously mentioned, these decoding strategies are utilized primarily in neural chatbot response generation.

For top-k sampling, we use $k = 40$; our motivation for choosing this value is that in Radford et al. (2019), the authors use top-k sampling with $k = 40$ for the task of "conditional" (prompted) generation and achieve quite good results.[8] Nonetheless,

[6] Implementation: https://huggingface.co/transformers/model_doc/GPT-2.html

[7] The entire model and dataset must be loaded onto each GPU, so the number of GPUs does not matter. It is necessary to drastically limit the length of training and output sequences, and this would interfere with one of our primary axes of experimentation: response length.

[8] Example generated responses are located in Radford et al.

Fold	Size	Tokens Per-Example	Total Tokens
Train	Small	92.9 ± 82.8	21.4M
	Medium	206.0 ± 128.2	47.5M
	Large	718.4 ± 458.9	165.8M
Valid	Small	92.9 ± 80.2	1.2M
	Medium	206.1 ± 128.3	2.8M
	Large	714.4 ± 463.3	9.5M
Test	Small	91.4 ± 79.4	1.2M
	Medium	204.7 ± 124.1	2.6M
	Large	720.4 ± 455.9	9.3M

Table 1: Corpus sizes for each fold and response length. **Tokens Per-Example** indicates the mean number of tokens per-prompt/response pair ($\pm$ standard deviation). **Total Tokens** indicates the number of tokens in the entire corpus.

the problem of choosing the best k for one's task is present here, as described in Holtzman et al. (2020).

The majority of our investigation is focused on nucleus sampling, as it generally outputs higher-quality responses (Holtzman et al., 2020; Ippolito et al., 2020). The original nucleus sampling paper uses $p = 0.95$; we perform an ablation over values of p here to discover which value best suits this task, both with and without diverse decoding. For completeness, we use a wide range of p-values of $0.3, 0.5, 0.7, 0.9$, and include greedy search (represented by $p = 0$) and fully random sampling (represented by $p = 1$).

We apply diverse decoding objectives to evaluate how the diversity objective affects narrative generation. Maximum mutual information (MMI, Li et al. 2016a) is an objective function which promotes more diverse responses in the neural response generation task. This mitigates the "I don't know" problem in which all responses tend to converge to some high-probability sequence with no real content conveyed in response to the input sequence. Specifically, we implement the MMI antiLM (anti-language model) objective for GPT-2. Li et al. (2016a) find through fine-tuning over λ (the strength of the modification to the original decoding objective) that the best value is in the range of $\lambda = 0.1$ to $\lambda = 0.2$. We perform a small ablation over λ values as well, testing the values $0.2, 0.5, 0.8$; $\lambda = 0$ represents not using MMI decoding. We expect these diversity objectives to hurt our model's coherence in general, and that higher values of λ will exaggerate this effect. However, we also expect this to improve the interestingness and creativity of the generated stories.

Note that the MMI objective may be applied to any of the previous decoding methods. Thus, we have two axes of decoding experimentation: (1) sampling methods (top-k and nucleus sampling) and (2) diversity-promoting methods (MMI), each combination of which we test.

3.4 Evaluation

A combination of automatic and human evaluation is used to evaluate the quality of the generated narratives. The qualities important for narrative generation are coherency, fluency, interestingness, and relevance. For automatic evaluation, we employ test perplexity, distinct-n (Li et al., 2016a), and a BERT-based similarity metric, Sentence-BERT (sent-BERT, Reimers and Gurevych 2019). The former acts as a proxy for fluency, and the latter two act as proxies for interestingness. To use sent-BERT as a diversity metric, we use cosine distance instead of cosine similarity. Our motivation for choosing these diversity metrics is from Tevet and Berant (2020), who identified distinct-n and sent-BERT as the best metrics to evaluate their two targeted types of diversity—diverse form and diverse content, respectively.

For human evaluation, we employ a binary scale for coherency, fluency, and interestingness (e.g., "Is this response fluent? Yes (1) or no (0)"). The former metrics measure the quality of a generated narrative without respect to the input prompt. Relevance is a metric we will employ to measure how well the response matches the prompt. For this, we use the human evaluations for the above metrics to

(2019)'s Appendix.

select a p-value to judge for relevance. We do not bother judging the responses from every p-value, because responses that have poor coherency and fluency would also have poor relevance scores; so we only wish to judge the relevance of the best performing decoding strategy. Relevance is judged on the same binary scale as the other metrics.

We also correlate model performance with the length of the responses. We expect that the models will display an inverse correlation between response length and the metrics of coherence and relevance. Conversely, we expect a positive correlation between response length and interestingness. Fluency should not be harmed or improved as output length is varied.

3.5 Baseline

We employ the fusion model, the previous state-of-the-art approach for narrative generation before transformer models, from Fan et al. (2018) as a baseline. The fusion model is essentially an ensemble of two Conv seq2seq models, where one is pre-trained on the data and is used to boost the second model. In Fan et al. (2018) this model was also trained on the WritingPrompts dataset and was evaluated using perplexity and human evaluations such as a task matching a response to a handful of prompts. We use the binary relevance scoring system instead of this prompt/response matching task.

4 Results

4.1 Quantitative Results

	Response Length		
Model	Small	Medium	Large
GPT-2 Small	30.52	23.74	15.64
GPT-2 Medium	**25.08**	**19.34**	**13.19**
Fusion Model	44.20	39.03	34.71

Table 2: Perplexities of the GPT-2 models and baseline model after fine-tuning on WritingPrompts dataset with different response lengths. The fusion model is the baseline from Fan et al. (2018).

First, we note the perplexities of each model on each narrative length, shown in Table 2. GPT-2 Medium has the lowest perplexity within each dataset size, with GPT-2 Small having a fairly close perplexity despite having significantly fewer parameters. Comparatively, the fusion model has a very high perplexity. In general, perplexity decreases as the length of the response increases, though numbers are not necessarily comparable across dataset sizes since this a per-word metric. Nonetheless, these results suggest that we should generally expect GPT-2 Medium to be marginally more fluent than GPT-2 Small, and that both of these will output far better English than the fusion model.

	Creative	Fluent	Coherent
GPT-2 Small	**1.00**	0.94	0.31
GPT-2 Medium	**1.00**	**1.00**	**0.63**
Fusion	0.60	0.88	0.00
Agreement	0.79	0.68	0.32

Table 3: Human evaluations across model sizes and architectures on the medium dataset with $p = 0.9$. The same 8 samples were used for each model.

We find this to be true; see Table 3. The small model generally outputs interesting and fluent stories which are not coherent, and the medium model performs similarly. However, the medium model corrects some of the errors made by the small model, and this is noted in its relative increase in fluency and coherence. Meanwhile, the fusion model outputs stories which are generally fluent English grammatically, but completely meaningless semantically. Indeed, there are drastic topic shifts within sentences for the baseline, to the point that each sentence is so incoherent so as to be hard to read—and thus, uninteresting. See Section 4.2 for a more qualitative discussion of these findings.

	Creative	Fluent	Coherent
Nucleus	**1.00**	**0.94**	0.31
Top-k	0.69	0.63	**0.75**
Agreement	0.76	0.45	0.22

Table 4: Human evaluations across decoding methods. These results are for GPT-2 Small with medium-length responses. Agreement is measured using Fleiss' kappa between three annotators.

Next, we compare two decoding methods: top-k and nucleus sampling. We perform this comparison through a human evaluation over medium-length outputs using GPT-2 Small; see Table 4 for these results. We see that nucleus sampling tends to produce stories that are creative and fluent, but not always coherent. Top-k demonstrates the oppo-

site trends: it produces stories which are creative (but less often than nucleus sampling), sometimes yields broken English or completely degenerate output, and it manages to stay on-topic more often. These trends are consistent across model sizes, though we note that outputs tend to be more fluent and cohesive on average with GPT-2 Medium. As nucleus sampling seems to output higher-quality narratives on average, we focus on this decoding method in our remaining analyses on the effect of the p-value in nucleus sampling across model sizes, dataset sizes, and antiLM diverse decoding.

Next, we discuss the relationship between model size and the diversity of outputs. Table 5 contains dist-n and sent-BERT scores for all model sizes, p values in nucleus sampling, and response lengths. In general, as p increases, so does the diversity, as shown by the higher dist-n and sent-BERT scores with $p = 1$. This trend makes sense since $p = 1$ gives the model access to the entire set of tokens for sampling, whereas the lower p values only give access to a constricted token set. The only outlier is greedy decoding (i.e. "$p = 0$"), where the sent-BERT metric identifies the responses occasionally as more diverse than any other p value. We think this is due to the degeneracy of the greedy responses. As seen in the output examples in Table 8, greedy decoding can produce degenerate and repetitive responses, however the responses are each degenerate in different ways and with different words, thus leading to a higher sent-BERT score (which judges content diversity per response and is not a "bag-of-words" token count like dist-n).

As for the model comparison, for any given p value and response length, GPT-2 Medium tends to use a slightly larger variety of tokens per-response than GPT-2 Small. Meanwhile, the diversity of the fusion model outputs is quite low in comparison—typically due to the degeneracy of the output. Note also that the dist-n scores are the same for the medium and large response lengths; this is also due to the degeneracy of the output and the surprisingly short stories generated, even when trained on large data and when allowed to generate up to 1,000 tokens.

Also for dataset size comparison, the responses for the medium dataset are generally more diverse than the small dataset, but the diversity goes back down with the large dataset. We think this is due to the length of the generated responses, even though we allow the models trained on the large dataset

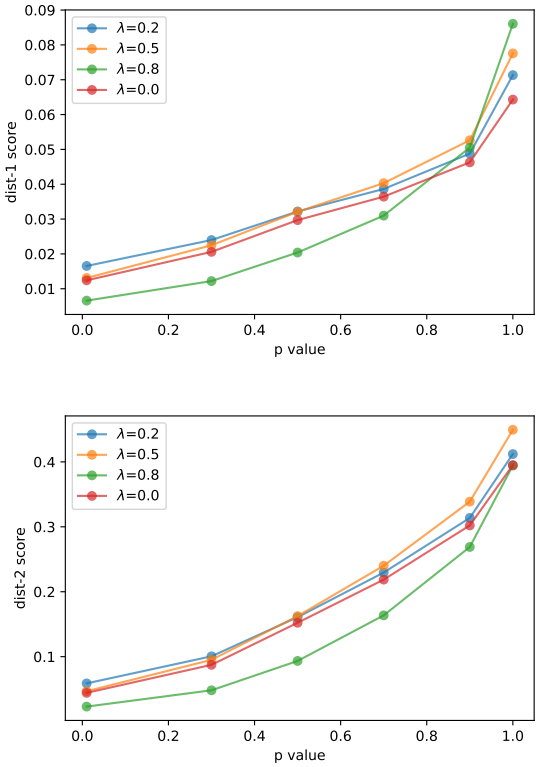

Figure 1: Plots comparing dist-1 (top) and dist-2 (bottom) scores across p values and MMI antiLM λ values. Note: $p = 0$ refers to greedy search.

to generate longer responses, they usually did not. Another explanation is that after a certain length the responses would start repeating information, thus causing lower diversity scores. Overall the medium-generated responses had the most diverse output.

Thus, we conclude that GPT-2 Medium produces the most diverse output, and we use GPT-2 Medium with the medium dataset for the human evaluation.

Do distinct-n or sent-BERT correlate with human judgments of quality? Not necessarily; see Table 6. Distinct-n increases monotonically with p, but we find that the best value of p is not the highest possible value it can take. Rather, we find that $p = 0.9$ tends to produce the best-quality output *on average*. Coherence tends to negatively correlate with p, while creativity positively correlates with p. Fluency is generally unaffected for reasonable values of p, though we note that completely random sampling and greedy search both output less fluent narratives.[9]

[9]Random sampling is less fluent due to the random series of tokens in the outputs, whereas greedy sampling is less fluent due to the degeneracy of the output.

		Small			Medium			Large		
Model	**Decoding**	**Dist-1**	**Dist-2**	**sent-BERT**	**Dist-1**	**Dist-2**	**sent-BERT**	**Dist-1**	**Dist-2**	**sent-BERT**
GPT-2 Small	greedy	0.014	0.048	**0.660**	0.012	0.043	0.614	0.002	0.008	**0.694**
	$p = 0.3$	0.026	0.104	0.610	0.018	0.075	0.644	0.004	0.018	0.670
	$p = 0.5$	0.037	0.165	0.607	0.027	0.137	0.659	0.006	0.035	0.669
	$p = 0.7$	0.045	0.224	0.625	0.034	0.206	0.668	0.011	0.079	0.670
	$p = 0.9$	0.057	0.299	0.641	0.044	0.291	0.672	0.021	0.193	0.673
	$p = 1.0$	**0.076**	**0.367**	0.650	**0.066**	**0.388**	**0.674**	**0.043**	**0.333**	0.674
GPT-2 Medium	greedy	0.021	0.072	0.600	0.014	0.050	0.634	0.002	0.008	**0.699**
	$p = 0.3$	0.031	0.120	0.605	0.021	0.087	0.655	0.006	0.026	0.578
	$p = 0.5$	0.041	0.176	0.615	0.030	0.152	0.664	0.010	0.048	0.586
	$p = 0.7$	0.048	0.232	0.629	0.036	0.219	0.636	0.016	0.105	0.593
	$p = 0.9$	0.059	0.301	0.642	0.046	0.302	0.676	0.030	0.232	0.598
	$p = 1.0$	**0.078**	**0.367**	**0.649**	**0.064**	**0.395**	**0.709**	**0.055**	**0.369**	0.600
Fusion Model	greedy	0.002	0.011	**0.791**	0.000	0.005	**0.788**	0.000	0.005	**0.788**
	$p = 0.3$	0.003	0.017	0.768	0.001	0.009	0.767	0.001	0.009	0.767
	$p = 0.5$	0.006	0.047	0.746	0.003	0.027	0.744	0.003	0.027	0.744
	$p = 0.7$	0.009	0.095	0.702	0.005	0.063	0.694	0.005	0.063	0.694
	$p = 0.9$	0.015	0.184	0.664	0.009	0.138	0.650	0.009	0.138	0.650
	$p = 1.0$	**0.031**	**0.292**	0.631	**0.020**	**0.244**	**0.616**	**0.020**	**0.244**	0.616

Table 5: Automatic diversity evaluations across models and decoding methods for each response length. The decoding methods represent a parameter sweep over the p value in nucleus sampling, where $p = 1$ corresponds to completely random sampling. The fusion model is a baseline from Fan et al. (2018).

	Creative	Fluent	Coherent
greedy	0.00	0.54	0.58
$p = 0.3$	0.21	0.67	0.42
$p = 0.5$	0.33	0.75	**0.67**
$p = 0.7$	0.58	**1.00**	**0.67**
$p = 0.9$	**1.00**	**1.00**	0.63
$p = 1.0$	**1.00**	0.71	0.13
Agreement	0.75	0.68	0.61

Table 6: Human evaluations over a sweep of p for nucleus sampling. Judgments over medium-length responses using GPT-2 Medium. For $p = 0.9$, the average relevance of the responses to their respective prompts (as judged by the same annotators) was 0.67.

	Creative	Fluent	Coherent
$\lambda = 0$	**1.00**	**1.00**	0.63
$\lambda = 0.2$	0.83	0.88	**0.71**
$\lambda = 0.5$	**1.00**	0.88	0.46
$\lambda = 0.8$	0.96	0.58	0.42
Agreement	0.57	0.76	0.22

Table 7: Human evaluations on medium-length outputs using GPT-2 Medium and $p = 0.9$ for various settings of λ in the MMI antiLM objective.

Finally, we analyze the effect of the MMI antiLM objective on narrative quality with various λ values. Figure 1 plots p vs. dist-n scores for each value of λ that we try. Surprisingly, there is not a consistent trend between values of λ and diversity. As we increase λ from 0.0 to 0.2 to 0.5, diversity increases slightly. However, as we increase λ further to 0.8, we see a sharp drop in distinct-n scores for $p < 0.7$, and a large increase for $p \geq 0.7$. More investigation will be necessary to determine the source of this unexpected drop.

Human evaluations across various diverse decoding strengths may be found in Table 7. Creativity does not seem to correlate well with λ, though we note that the diction is generally more interesting with higher values of λ. Fluency tended to decrease as λ increased, though this trend was inconsistent. Coherence was little-affected by varying λ. The fluency and coherence trends are mathematically sensible, since the MMI antiLM objective subtracts a language model from the conditional distribution $p(T|S)$, thus increasing the importance of the prompt and decreasing the importance of generating fluent English.

4.2 Qualitative Results

In this section, we analyze the quality of narratives by looking at the outputs. See Appendix A for generated narratives from a variety of model architectures, sizes, and decoding hyperparameters.

4.2.1 Top-k vs. Nucleus Sampling

For most reasonable settings of p, nucleus sampling tends to produce stories which are dramatic, vivid, and fun to read, but which do not often stay on topic. Indeed, the outputs demonstrate two main types of errors: (1) cramming too many topics into one story, and (2) sudden shifts in topic. Indeed, they often read like a new writer excited to put every science fiction trope into one short story that does not necessarily have a singular message. The English is impeccable and it is generally enjoyable to read—and even demonstrates some basic understanding of introductions vs. conclusions—but it does not have any significant semantic flow.

Top-k sampling, however, demonstrates quite extreme variance. Some of the generated stories feel almost human-like with how on-topic they remain for multiple paragraphs—but they are about safe and boring topics and generally employ very common token collocates, which makes the output feel uncreative and uninteresting. Other stories are dramatic, but almost dream-like due to the stream-of-consciousness incoherent flow. Yet other stories are completely unintelligible and show signs of neural text degeneration (Holtzman et al., 2020).

Note that the inter-annotator agreement per-metric varies quite markedly. Creativity seemed fairly easy to judge: if it was entertaining or semantically interesting, it was creative. The moderate agreement for fluency is somewhat surprising, but this is perhaps due to a matter of degree: a few typos might sway one annotator, while another might have a higher threshold for such errors. We remark that the variance of narrative quality correlates inversely with agreement. When there are more borderline-quality outputs, the agreement tends to be lower; When output is more consistently good or bad, agreement is much higher. This is why we tend to see higher agreements for nucleus sampling evaluations than top-k evaluations.

4.2.2 Nucleus Sampling

When p is high, we generally see more interesting and vivid narratives with good diction and fluency scores, but also stories which have no single cohesive plot. When p is low, we see more repetitive word choice but higher cohesion. However, when p is very low (e.g., $p = 0.3$ or lower), the output is degenerate. Generally, values of p around 0.7 output coherent but less-interesting stories, whereas $p = 0.9$ generally output interesting stories that were often cohesive. This is intuitive with how p is restricting the sampling space; when p is too small, too many options are removed and the model cannot generate fluent text, which then is also not coherent.hen And then when $p = 1$, this is random sampling and no tokens are removed, so the probability tail makes it more likely for the model to choose unlikely tokens, which makes it interesting, but also less fluent/coherent.

We found that $p = 0.9$ was the best setting for the best narratives in general, though this should be tuned based on one's desired balance of creativity and cohesion.

4.2.3 Diverse Decoding

For smaller values of λ, MMI had a surprisingly small effect on the output of the models. Within a given p value, increasing MMI values up to 0.5 seemed to result in slightly more interesting diction for the small models. However, increasing λ also resulted in slightly less grammatical English. Coherence seemed to be unaffected by changing values of λ.

More interesting is that the intensity and disturbingness of the subject matter tended to increase with λ. Indeed, subjects such as abuse, murder, as well as crude jokes featuring sexuality and dark humor, became more common as λ increased. This may not necessarily be a positive or negative trend; if one wishes to generate stories which are more intense and disturbing, and one's language model is sufficiently high-quality to take a small fluency hit, then this may be a worthwhile method to employ. Nonetheless, we do not have a clear mathematical explanation for this, since the MMI antiLM objective simply increases the importance of the prompt while decreasing the importance of the language model.

An unfortunate effect of higher λ values like 0.5 and 0.8 is that narratives tend to switch (seemingly randomly) between fluent English and complete degeneracy. The token "eleph" is often repeated, regardless of subject matter or sentence position. More investigation would be necessary to determine the source of this random degeneration effect.

4.2.4 Correlating Automatic Metrics with Quality

Thus far, we have discussed how perplexity, distinct-n, and sent-BERT vary with various model architectures/sizes, decoding approaches, and hyperparameters. However, what do these quantities say about the quality of generated narratives? In

general, we see the following trends: (1) Lower perplexity is always better. This correlates mainly with fluency and non-degenerate output. (2) Very low distinct-n scores indicate consistent neural text degeneration. (3) Very high distinct-n scores indicate non-cohesive (but fluent and creative) narratives. This is not to say that there is any one-size-fits-all distinct-n score; indeed, this is not a causal relationship. Rather, we find that this quantity can be a helpful heuristic when comparing across model configurations at a high level.

5 Conclusions

Our results suggest that GPT-2 generally outputs better narratives than the most recent non-GPT-based neural model. Additionally, the larger the model, the better. While GPT-2 Large may be infeasible for very long sequence generation, it is possible to use GPT-2 Medium for all narrative lengths generated here.

With respect to decoding methods, nucleus sampling generally outperformed top-k sampling. The former gives one more control over specific features of the output, such as coherence and creativity, as the p hyperparameter seems to correlate more strongly with these features than the k hyperparameter. We recommend keeping this hyperparameter to the following range: $0.7 \leq p < 1$, using higher values for more creative output and lower values for more cohesive output.

Diverse decoding did not necessarily have a consistent relationship with the automatic or human evaluation metrics employed herein, but it did correlate with qualitative judgments of story intensity. If one wishes to generate more disturbing or crude subject matter, then this method may be useful—especially since one may tune the intensity by tuning λ. However, the stories generated by nucleus sampling with higher values of p are already quite creative, so we do not believe that diverse decoding is necessary to achieve interesting output.

Output length did not significantly affect any of the metrics employed here except cohesion. This is unsurprising, as it is easier for a generated story to veer off-topic when it is allowed to generate more content. Generally, with longer output, we recommend lowering p to keep narratives more cohesive; this will need to be tuned to achieve the best balance of creativity and cohesion for one's desired narrative length.

Acknowledgments

The authors thank João "J-Wow" Sedoc, for designing and teaching the wonderful Deep Learning Methods for Automated Discourse course at Johns Hopkins. We also thank classmates who provided great feedback on our paper.

References

Kyunghyun Cho. 2016. Noisy parallel approximate decoding for conditional recurrent language model. *CoRR*, abs/1605.03835.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. *CoRR*, abs/1805.04833.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Daphne Ippolito, Joao Sedoc, Aaron Mueller, Alexandra DeLucia, and Lisa Li. 2020. Zoom meeting.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 839–849, San Diego, California. Association for Computational Linguistics.

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. 2018. Another diversity-promoting objective function for neural dialogue generation. *CoRR*, abs/1811.08100.

Alec Radford. 2018. Improving language understanding by generative pre-training.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP*.

Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. *ArXiv*, abs/2004.02990.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, Brussels, Belgium. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1815–1825, Red Hook, NY, USA. Curran Associates Inc.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724.

A Example Outputs

All examples start on the following page. We report narrative responses given a single prompt for various model architectures/sizes, decoding methods, and hyperparameter sweeps.

top-p	Prompt: “[WP] You just started a new job in a large office building and what you do n’t know is that at twelve o’clock every day , there is a office wide Purge style battle to the death , those who survive an hour get a significant pay raise and can come back tomorrow . A bell just went off , it ’s twelve o’clock .”
greedy	“ Hey, he
$p = 0.3$	I was nervous. <newline><newline>I was nervous about the job. <newline><newline>I was nervous about the pay.
$p = 0.5$	The first thing I noticed was the absence of people. I wasn’t sure why, but I couldn’t shake the feeling that it wasn’t just the people that were gone. I mean, it wasn’t like I hadn’t been here before. I mean, I was in the office for a month, but it felt like a year. <newline><newline>I wasn’t sure what was going on, but I could tell that something was wrong. <newline><newline>“ Hello? ”
$p = 0.7$	My boss was a loner. He liked to work late and I didn’t know why. I didn’t want to work late. I just didn’t have the patience for it. <newline><newline>I don’t know what happened. <newline><newline>“ You got ta be kidding me, ” he said, “ a sixteen year old would just walk into the office and run out of work? ”
$p = 0.9$	All four of the chimes clicked and the chandelier snapped, as the room dimmed and grew dim until the only light came from an overhead television, showing the “ The Purge ” live. In the background, a towering wall of reds, greens, and yellows flashed in contrast to the background colors, casting warm green shimmers across the television. <newline><newline>The chandelier struck the ceiling, crumpling under the weight. A small circle of lights in the ceiling blinked and illuminated what had been a desk, until two figures entered it. <newline><newline>The first man, dressed in a yellow and white baseball cap and sunglasses, stood tall, wearing khakis and a button-down shirt. A muscular build and hairy arms waved about as he spoke.
$p = 1.0$	“Chad’s Melatonin Impaired Heart implant had occurred during the Golden elimination pilot. “ Critically, ” the docs explained to the board, “ Third lapse, unlocking the minimum Carol Corporation-mandated for the Restricted Solid Species Program. We penalize those who have such lapses as early and late entry personnel. ” <newline><newline>Chad felt pain in the core of his neck - compressing saliva into his neckline, seeing his neighbors ’ streaks poking out, “ That doesn’t work, I still live in Seattle ’ s Milliner Square until May 17th, ” thought Chad as he fell down the empty stairs the hospital had installed that morning. Without warning, his head slammed into the steel sill beneath him and he felt a sharp pain begin to run down his spine. The pain apparently doomed him to eternity in a single meditation, the disease never getting better. <newline><newline>After five ER visits, nine Lab ultrasound tests, six minor surgeries, pressure checks, one heart-heated ultrasound, Chad came back for surgeries, six of which were removeable .

Table 8: Medium-length responses from GPT-2 Medium with a sweep of nucleus sampling parameter p.

Model	Prompt: “[WP] You just started a new job in a large office building and what you do n’t know is that at twelve o’clock every day , there is a office wide Purge style battle to the death , those who survive an hour get a significant pay raise and can come back tomorrow . A bell just went off , it ’s twelve o’clock .”
GPT-2 Small, $p = 0.9$	This reminds me of a famous episode in The Show where there was a real White noise, then it sounded like we were having a gun fight. There was screaming then explosions, when the “ Purge ” had finally ended, people screamed louder than anyone had expected. I’ve never witnessed a screaming from somebody shouting things like that. I’m not kidding. <newline><newline>The voices also were always recorded as late as they had been recorded, at sevenpm everyone was going into work, panicking. At sixpm it was over. People went into work then went in the closet, then out into the hallway. People went in as the seconds dragged on, from seven to twelve. By sixpm everyone was screaming again. I only noticed it out of habit later on as a storm slowly shifted in and out of place and went into a raucous, god forsaken night. It wasn’t till the next six that it became clear what was going on. There was another group of kids screaming. “ IT ’S GOIN’OVER AGAIN! ” They screamed then shouting again “ IT ’S GIVEN TO EVERYONE! ” They threw their arms around one another and they shouted in the same tone as their parents
GPT-2 Med., $p = 0.9$	All four of the chimes clicked and the chandelier snapped, as the room dimmed and grew dim until the only light came from an overhead television, showing the “ The Purge ” live. In the background, a towering wall of reds, greens, and yellows flashed in contrast to the background colors, casting warm green shimmers across the television. <newline><newline>The chandelier struck the ceiling, crumpling under the weight. A small circle of lights in the ceiling blinked and illuminated what had been a desk, until two figures entered it. <newline><newline>The first man, dressed in a yellow and white baseball cap and sunglasses, stood tall, wearing khakis and a button-down shirt. A muscular build and hairy arms waved about as he spoke.
Fusion Model, $p = 0.9$	He was n’t ready to die . I did . The name on his chest would kill him if he wanted to know . They wanted me to accept him as he wanted me to be not ready . I did it for him but he loved me . <newline><newline>I was the one that started a revolution . It was a simple matter of encouraging him and all the loyal soldiers that murdered him . And yet I never seemed to know what he had to do to deserve my right to be born and robbed . If only I knew . I could feel the faintest drop of power beneath my footsteps as he died . I felt the need to put some strings around him . But he had so much to offer . He seemed to feel I was the leader of the group and I was ready . He taught me the fact that he was supposed to be the most hated man in the country . So the world slowed . <newline><newline>And so he fought me . I was ready . He lost every step of his very being when he

Table 9: Medium-length responses from GPT-2 Small, GPT-2 Medium, and Fusion (baseline) models with nucleus sampling parameter $p = 0.9$. We also include the sweep of MMI-antiLM λ values for direct comparison with using the diverse decoding objective.

Model, Decoding	Prompt: "[WP] You just started a new job in a large office building and what you do n't know is that at twelve o'clock every day , there is a office wide Purge style battle to the death , those who survive an hour get a significant pay raise and can come back tomorrow . A bell just went off , it 's twelve o'clock ."
GPT-2 Med., $p = 0.9$	All four of the chimes clicked and the chandelier snapped, as the room dimmed and grew dim until the only light came from an overhead television, showing the " The Purge " live. In the background, a towering wall of reds, greens, and yellows flashed in contrast to the background colors, casting warm green shimmers across the television. <newline><newline>The chandelier struck the ceiling, crumpling under the weight. A small circle of lights in the ceiling blinked and illuminated what had been a desk, until two figures entered it. <newline><newline>The first man, dressed in a yellow and white baseball cap and sunglasses, stood tall, wearing khakis and a button-down shirt. A muscular build and hairy arms waved about as he spoke.
GPT-2 Med., $k = 40$	The first time it happened I remember it vividly. The first time I met it. We got our first coffee, we sat on the train and talked and made plans but when I first saw it coming I knew something was going to happen. There was a man in white with a clipboard. I said " how you doing? " and he asked " what the hell did I just read you about ". I told him nothing he went through what the hell I knew everything about the Purge and I knew that I had to be careful because if I wasn't careful it would happen again. He looked at me expectantly and I knew immediately that I had to go and try to help people ' s lives. No one else would go so I did as I was told. <newline><newline>He asked me if I had time for something more serious than just reading
Fusion Model, $p = 0.9$	He was n't ready to die . I did . The name on his chest would kill him if he wanted to know . They wanted me to accept him as he wanted me to be not ready . I did it for him but he loved me . <newline><newline>I was the one that started a revolution . It was a simple matter of encouraging him and all the loyal soldiers that murdered him . And yet I never seemed to know what he had to do to deserve my right to be born and robbed . If only I knew . I could feel the faintest drop of power beneath my footsteps as he died . I felt the need to put some strings around him . But he had so much to offer . He seemed to feel I was the leader of the group and I was ready . He taught me the fact that he was supposed to be the most hated man in the country . So the world slowed . <newline><newline>And so he fought me . I was ready . He lost every step of his very being when he
Fusion Model, $k = 40$	It was a sunny Monday morning when I woke up to the noise of my alarm going off . I got up from my bed , got out of bed , and went into the bathroom and took off my coat . It was n't exactly a normal morning . I walked into the bathroom and put on my shoes , and put on some pants , and went to the bathroom . The light from the bathroom was n't going to change anything . I walked out of the bathroom and went to the bathroom . It was a good morning . My morning routine was going well in bed , and I was going to see some shit , so it was good . <newline>I went to the bathroom . It was the first step in my morning shift , so I took off my pants and

Table 10: Medium-length responses from GPT-2 Medium and the Fusion (baseline) model with top-k and nucleus sampling.

λ	Prompt: “Your baby starts crying every time you leave their room . Finally , you pick up your child and leave the room together . As soon as you step out, you hear crying coming from the room .”
$\lambda = 0.0$	“ Remember that milk, honey? ” “ ... Yeah? ” I gestured toward my old neighbor who had come in to see me having a late night. “ Oh my god, honey...”
$\lambda = 0.2$	“ I know mom, but I have to go. ” My wife gave a confused glance at my place and said, “BABY!” I carefully set the baby down on the park bench next to my house. I take the seat down in the bushes. Here I would meet her, and it would be over.
$\lambda = 0.8$	I could have called the police, called the family or told the therapist. But I don’t have time to get to those things. I must now face the reality. It is too late. I am the bastard who left the baby inside the black vacuum box. The one in the bus seat. I am the one who left the toddler outside. I am the one who left the couple with a red cupcake on the park bench.

Table 11: Medium-length stories generated using GPT-2 Medium with nucleus sampling ($p = 0.9$) and various diverse decoding strengths λ.