

Decoding Strategies for Interactive Narrative Generation

Alexandra DeLucia*
aadelucia@jhu.edu

Aaron Mueller*
amueller@jhu.edu

Xiang “Lisa” Li
xli150@jhu.edu

João “JWow” Sedoc
joao.sedoc@gmail.com



JOHNS HOPKINS
UNIVERSITY

*Equal contributors to project epicness

Personal Introduction



PhD student under Mark Dredze in the
Center for Language and Speech Processing at
Johns Hopkins University

Primary research: Civil unrest detection/forecasting
with Twitter

Secondary “research”: Laughing at 🤖 chatbots say



[@Alexir563](https://twitter.com/Alexir563)



alexandradelucia.com



aadelucia@jhu.edu

³ What is interactive narrative generation?

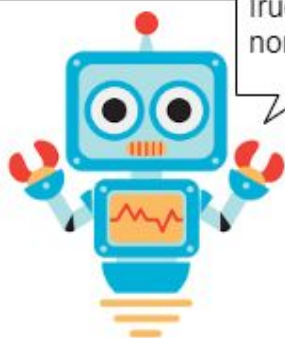
In 2020 there was a virus that some considered to be the apocalypse .



“ What do we do? ” asked Irudim as he gestured to the two million persons in attendance, tired and shaken.

“ Yeah, we die, ” interjected Yuzsagm, “ We get sick. ”

Irudim opened his mouth, but his words did not catch, nor did the rest of the tal...



Motivation

Research in neural
response generation
decoding

Research in narrative
generation decoding



Now kiss

Background

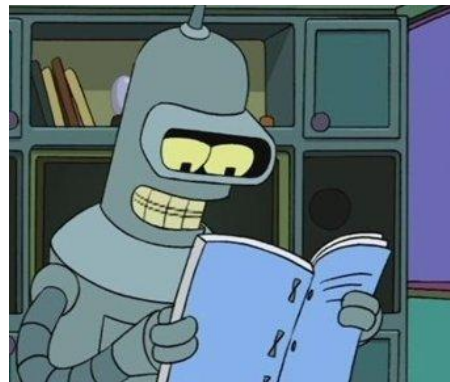
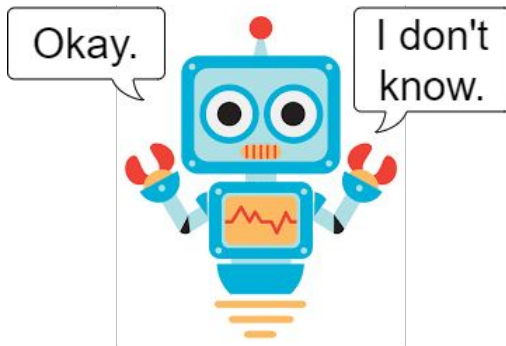
Decoding

Diverse Decoding
(The “I don’t know”
problem)

Other narrative
generation models

This little piggy went to the **market**

history generated token



Background

Decoding Methods

What is **decoding**?

Language modeling: how the model picks the next word in the sequence

$$P(w_i | w_1, w_2, \dots, w_{i-1})$$

Focus on **sampling-based** decoding methods

This little piggy went to the ____

Token	Probability
market	0.5
farm	0.3
bathroom	0.1
airport	0.05
Bloomingdales	0.05

Background

Decoding Methods

Common sampling decoding methods

This little piggy went to the ____

Token	Probability
market	0.5
farm	0.3
bathroom	0.1
airport	0.05
Bloomingdales	0.05

Background

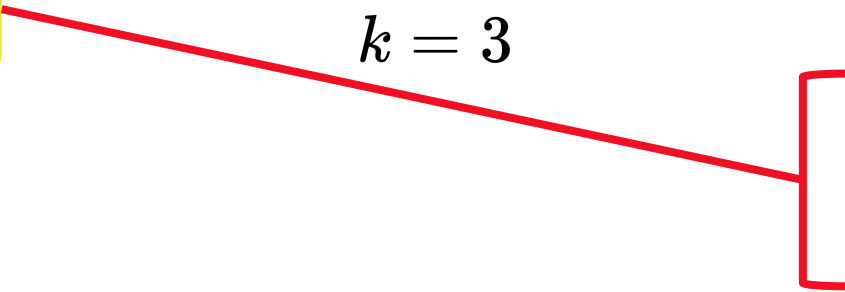
Decoding Methods

Common sampling decoding methods

Top-k
sampling

$$k = 3$$

This little piggy went to the ____



Token	Probability
market	0.5
farm	0.3
bathroom	0.1
airport	0.05
Bloomingdales	0.05

Background

Decoding Methods

Common sampling decoding methods

Top-k
sampling

Nucleus
sampling
(top-p)

$$p = 0.9$$

This little piggy went to the ____

Token	Probability
market	0.5
farm	0.3
bathroom	0.1
airport	0.05
Bloomingdales	0.05

Background

Decoding Methods

Common sampling decoding methods

Top-k
sampling

Nucleus
sampling
(top-p)

Random
sampling

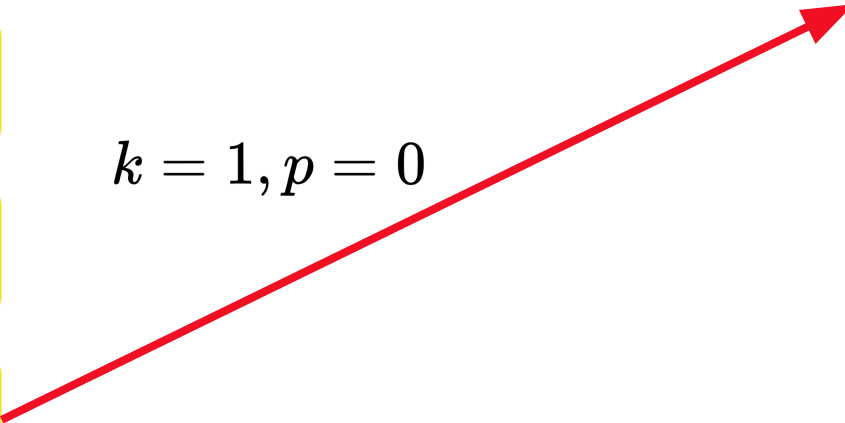
$$k = \#tokens, p = 1.0$$

This little piggy went to the ____

Token	Probability
market	0.5
farm	0.3
bathroom	0.1
airport	0.05
Bloomingdales	0.05

Common sampling decoding methods

Top-k
samplingNucleus
sampling
(top-p)Random
samplingGreedy
(argmax)

$$k = 1, p = 0$$


This little piggy went to the ____

Token	Probability
market	0.5
farm	0.3
bathroom	0.1
airport	0.05
Bloomingdales	0.05

Background

Diverse Decoding

Common issue with decoding:

Generic responses have high probability

Because generic responses:

- are more common than others
- can fit in many situations

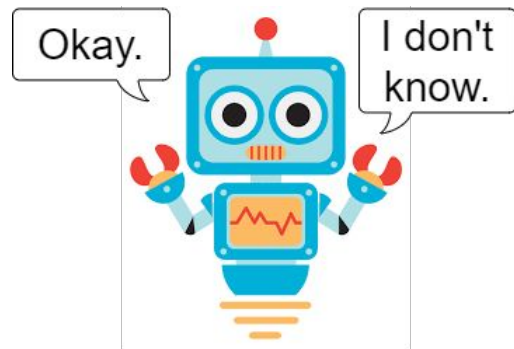
Sure.

Great.

I'm sorry.

Nothing.

The “I don’t know” Problem



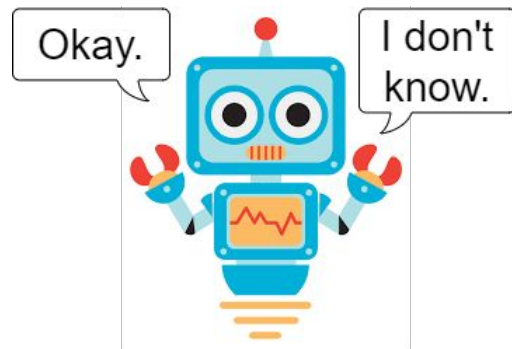
Background

Diverse Decoding

One possible solution:

Use an objective function that maximizes the mutual information (MMI) between the **source** and **target**

The “I don’t know” Problem



One possible solution:

Use an objective function that maximizes the mutual information (MMI) between the **source** and **target**

MMI-antiLM model (anti language model)¹

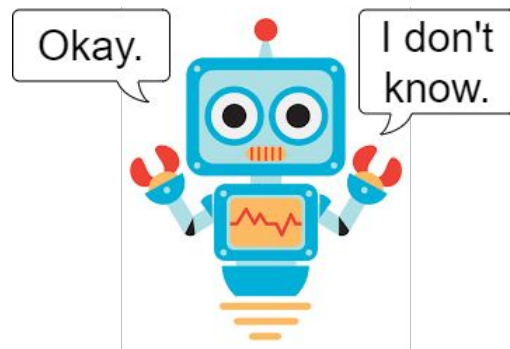
$$\hat{T} = \arg \max_T \{ \underbrace{\log p(T|S)}_{\text{Conditional language model}} - \underbrace{\lambda \log p(T)}_{\text{weight subtract language model}} \}$$

Conditional language
model

weight

subtract language model

The “I don’t know” Problem

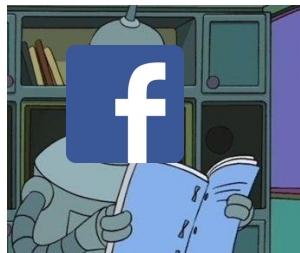


1. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages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

Background

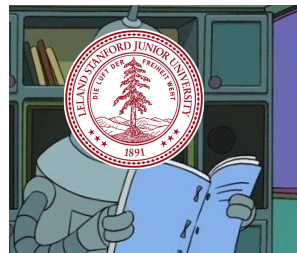
Other Narrative Generation Models

Fusion model



- Introduced in Fan et al.¹
- Ensemble of two convolutional seq2seq models
- Created WritingPrompts dataset from subreddit r/WritingPrompts
- Trained Fusion model on WritingPrompts

Fine-tuned GPT-2 (Small)



- Introduced in See et al.²
- Trained GPT-2 Small on WritingPrompts
- Compared model performance to Fusion model
- Performed parameter sweep for top-k decoding

1. Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. CoRR, abs/1805.04833.
2. See, Abigail et al. "Do Massively Pretrained Language Models Make Better Storytellers?" CoNLL (2019).

Research Questions

How do existing decoding methods from neural response generation perform in narrative generation?

1. How do different neural architectures and model sizes affect quality?
2. Top-k vs. nucleus sampling?
3. **Different thresholds p for nucleus sampling?** 
4. **Diverse decoding (MMI antiLM) with different strengths λ ?** 
5. How do the above change with respect to different length narratives?

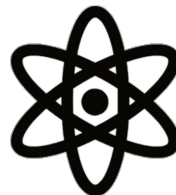
**GPT-2
Medium**

Methods

Fine-tune GPT-2 for
narrative generation



Decoder parameter
sweep

 λ 

Evaluate



amazon mechanical turk

Methods

Fine-tune GPT-2 for narrative generation

- Hugging Face implementation of GPT-2 Medium^{1,2} 🤗
- Fine-tuned on WritingPrompts
 - Collection of prompts and responses from subreddit r/WritingPrompts collected by Fan et al³



Posted by u/Prowlerbaseball 2 years ago 📄

Writing Prompt [WP] All of the "#1 Dad" mugs in the world change to show the actual ranking of Dads suddenly.

```
<|startoftext|>[WP] All of the "#1 Dad" mugs in the world change to show the actual rankings of Dads suddenly [RESPONSE] I entered the kitchen to witness an event I had never seen before... my Father with his head in his hands.<|endoftext|>
```

1. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners
2. https://huggingface.co/transformers/model_doc/GPT-2.html
3. Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. CoRR, abs/1805.04833.

Methods

Fine-tune GPT-2 for narrative generation

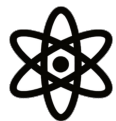
Create three different sized response sets based on WritingPrompts dataset



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

Methods

Decoder parameter sweep



Nucleus sampling

$$p = 0.0 \leftarrow \text{greedy}$$

$$p = 0.3$$

$$p = 0.5$$

$$p = 0.7$$

$$p = 0.9$$

$$p = 1.0 \leftarrow \text{random sampling}$$

expect best
performance



expect direct
correlation
with
degeneracy



MMI-antiLM

 λ

$$\lambda = 0.0 \leftarrow \text{No MMI}$$

$$\lambda = 0.2$$

$$\lambda = 0.5$$

$$\lambda = 0.8$$

How “**diverse**” is the model output?

Tevet and Berant¹ identify two types of diversity and the best way to measure each one:

Diversity of **content** and **form**

what to say how to say it

- ✓ Diverse form
- ✓ Diverse content

- ✓ Diverse form
- ✗ Diverse content

Question: So what did I miss in the first 20 minutes?

Set A

- Pretty much everything.
- Nothing, really.
- You won't believe what happened!
- Why do you even care?
- What were you doing that was more important than this?

Set B

- Not much.
- It was pretty dull.
- Blah, you didn't miss anything.
- Not anything that important.
- Very little, it was uneventful.

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

How “**diverse**” is the model output?

Tevet and Berant¹ identify two types of diversity and the best way to measure each one:

Diversity of **content** and **form**



✓ Diverse form
✓ Diverse content

✓ Diverse form
✗ Diverse content

Question: So what did I miss in the first 20 minutes?

Set A

- Pretty much everything.
- Nothing, really.
- You won't believe what happened!
- Why do you even care?
- What were you doing that was more important than this?

Set B

- Not much.
- It was pretty dull.
- Blah, you didn't miss anything.
- Not anything that important.
- Very little, it was uneventful.

1. Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. ArXiv, abs/2004.02990.
2. Nils Reimers and Iryna Gurevych. 2019. SentenceBERT: Sentence embeddings using siamese bertnetworks. In EMNLP/IJCNLP.
3. 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, Cali

Evaluation

Human

Interesting

Is the story fun
to read?

Coherent

Do the sentences
form an
understandable
plot?

Fluent

Is the story written
in grammatical
English?

Relevant

Does the story
follow the prompt?

Evaluation

Human

Interesting

Is the story fun
to read?

UNDER CONSTRUCTION!

Is the story
easy to understand
plot?

Is the story written
in grammatical
English?

Relevant

Does the story
follow the prompt?

greedy

Hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey,
hey,
hey,
hey,
hey,
hey,
hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey, hey,

degenerate

I was nervous.

$p = 0.3$

I was nervous about the job.

Still repetitive

I was nervous about the pay.

Results

Examples: nucleus sampling sweep

[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 .

$$p = 0.5$$

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. I wasn't sure what was going on, but I could tell that something was wrong. `` Hello? "

good

$$p = 0.7$$

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.

I don't know what happened.

good

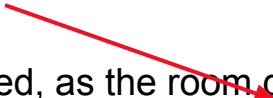
`` You got ta be kidding me, " he said, `` a sixteen year old would just walk into the office and run out of work? "

Results

Examples: nucleus sampling sweep

[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 .

wut



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.

$$p = 0.9$$

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.

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.

Results

Examples: nucleus sampling sweep

[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 .

``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. "

spicy

$$p = 1.0$$

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.

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 .

Results

Automatic: nucleus sampling sweep

$\uparrow p = \uparrow \text{diversity}$

degenerate 

Nucleus Sampling	Distinct-1	Distinct-2	Sent-BERT
greedy	0.014	0.050	0.634
$p = 0.3$	0.021	0.087	0.655
$p = 0.5$	0.030	0.152	0.664
$p = 0.7$	0.036	0.219	0.636
$p = 0.9$	0.046	0.302	0.676
$p = 1.0$	0.064	0.395	0.709

Simple trend does not accurately capture the performance from the examples:

Need human evaluations

Results

Examples: diverse decoding sweep
with $p=0.9$

[WP] 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.

wut

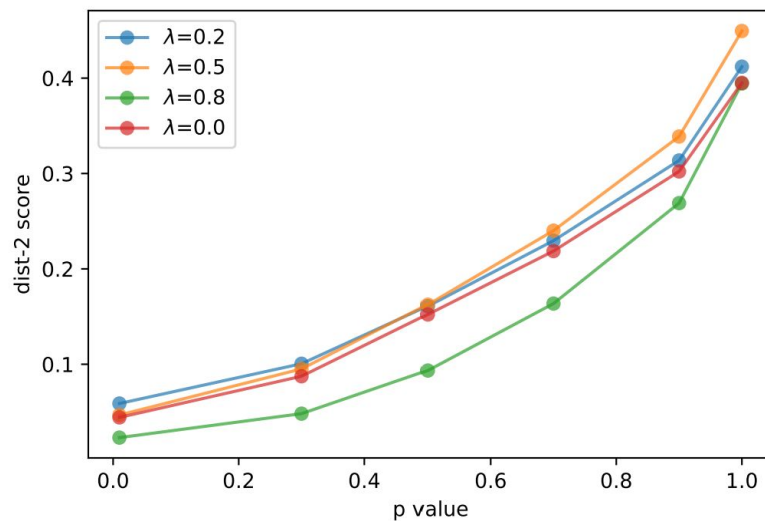
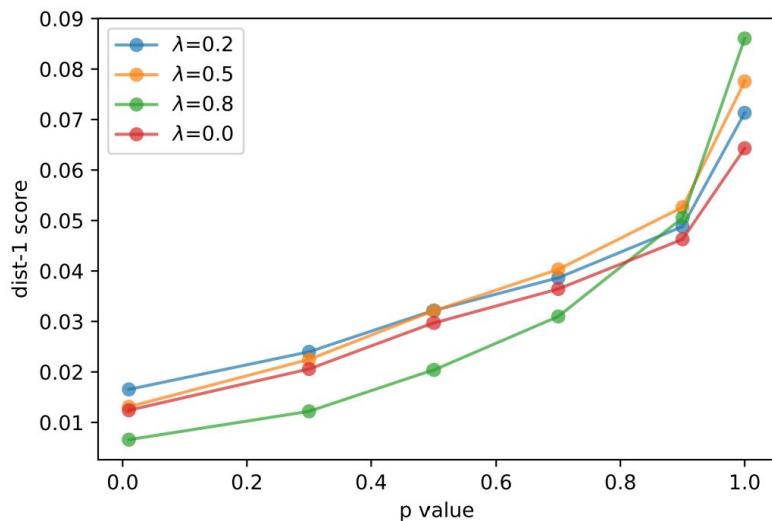
$\lambda = 0.8$

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.

Results

Automatic: diverse decoding sweep

- Form diversity increases as nucleus sampling increases
- Mostly no gains with a diverse decoding parameter over 0.2
 - According to automatic results



Summary

Methods

- Fine-tuned GPT-2 Medium for narrative generation with prompts and responses from subreddit r/WritingPrompts
- Performed nucleus sampling decoding parameter sweep
- Performed diverse decoding parameter sweep for MMI-antiLM objective
- Evaluated diversity of output with distinct-n and sent-BERT

Findings

- Nucleus sampling p-value between 0.7 and 0.9 work best, with lower values being degenerate
- Diverse decoding hyperparameter is directly correlated with response diversity
- There is a tradeoff between diversity and coherence

Thank you for your time

Check out our bot on reddit:

[u/StoryBotAI](https://www.reddit.com/u/StoryBotAI)

Join the bot narrative
generation community:

[r/StoryGenAI](https://www.reddit.com/r/StoryGenAI)



[@Alexir563](https://twitter.com/Alexir563)



alexandradelucia.com



aadelucia@jhu.edu

[RESPONSE] Alexandra simply went about her day. She went to school, played tennis, watched television, did some personal training, did some yoga. She lived in the old house she lived in with her parents and sisters, with no other family. Sh
(7.7s to generate response)

Prompt > Lisa

[RESPONSE] I could feel the water trickling out of the drain by the garage door. I was pretty damn thirsty, and I could feel the pile of cash and junk on the floor. I grabbed a beer from the fridge and climbed the stairs and
(7.7s to generate response)

Prompt > Aaron

[RESPONSE] It's been 22 years. <newline> <newline> It's been 22 years. <newline> <newline> It's been 22 years since my grandfather died in a car crash, laying in his
(7.7s to generate response)

Extras

Results

Automatic

Model	Decoding	Small			Medium			Large		
		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

Results

Examples: Fusion

[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 .

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 .

$$p = 0.9$$

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 .

And so he fought me . I was ready . He lost every step of his very being when he

Further Reading

Narrative Generation

- Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. CoRR, abs/1805.04833.
- See, Abigail et al. “Do Massively Pretrained Language Models Make Better Storytellers?” CoNLL (2019).
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners

Further Reading

Decoding and Diverse Decoding

- 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.
- 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.
- Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. ArXiv, abs/2004.02990.
- 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, Cali
- Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. 2018. Another diversity promoting objective function for neural dialogue generation. CoRR, abs/1811.08100
- Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. CoRR, abs/1611.08562.

#f10f24

#547d8f

#f2e436

#f4ad2c

#f17427

