The Curious Case of Neural Text *De*generation

Ari Holtzman Jan Buys Maxwell Forbes Yejin Choi

Reporter : Xiachong Feng

Author

- **Ari Holtzman**
- PhD student at the University of Washington
- Advised by Yejin Choi

Publications

Learning to Write with Cooperative Discriminators
Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi
In *Proceedings of the Association for Computational Linguistic* (ACL), 2018

*De*generation

Context:

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Continuation (BeamSearch, b=10):

"The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the…

Randomization over Maximization?

- 科学家们有个令人震惊的发现，在安第斯山脉一个偏远且没被开发过的山谷里，生活着一群独角兽。更加让人讶异的是，这些独角兽说着完美的英文。
- 这些生物有着独特的角，科学家们就以此为它们命名，叫Ovid’s Unicorn。长着四只角的银白色生物，在这之前并不为科学界所知。
- ……
- 虽然，这些生物的起源还不清楚，但有些人相信，它们是一个人和一个独角兽相交而诞生的，那时人类文明还不存在。Pérez教授说：“在南美洲，这样的现象很常见。”
- ……
- 如果要确认它们是消失种族的后裔，DNA检测可能是唯一的方法。

Randomization over Maximization?

- High-quality article about Ovid's Unicorn, written by **GPT-2**
- **Randomness** in the decoding method
- ***Top*-k sampling** that samples the next word from the top k most probable choices
- Instead of aiming to decode text that ***maximizes* likelihood**. (Beam search)

Task

- **Open-ended Generation**
- Given a sequence of *m* tokens *x1 . . . xm* as context, the task of open-ended language generation is to generate the next ***n* continuation tokens** to obtain the completed sequence
- **Conditional story generation** and **contextual text continuation**
- **Dataset**: WritingPrompts
  - Each example consists of a context of 5 sentences with a maximum of 200 tokens; the task is to continue the text by generating the 200 next tokens (the continuation).
- Language Model :GPT-2

Non-open-ended Generation

- Text generation tasks are defined through (input, output) pairs, such that the output is a close ***transformation*** of the input.
- **Machine translation, data-to-text generation, and summarization.**

Questions

Why does decoding with **beam search** from a strong language model lead to such **degenerate** text?

Why does sampling from a truncated vocabulary distribution perform better than sampling from the whole distribution?

What is the most principled method of truncation currently available?

Why Does Probability Maximization Lead to Degenerate Text?

- Decoding strategies which assume that the model assigns **higher probability to higher quality text**, and therefore aim to find the output with the highest likelihood.

$$x_{m+1:n} = \underset{x_{m+1:n}}{\operatorname{argmax}} P(x_{m+1:n} | x_{1:m})$$

- *Beam search*
- *Greedy* decoding (beam = 1)

Beam search

Human

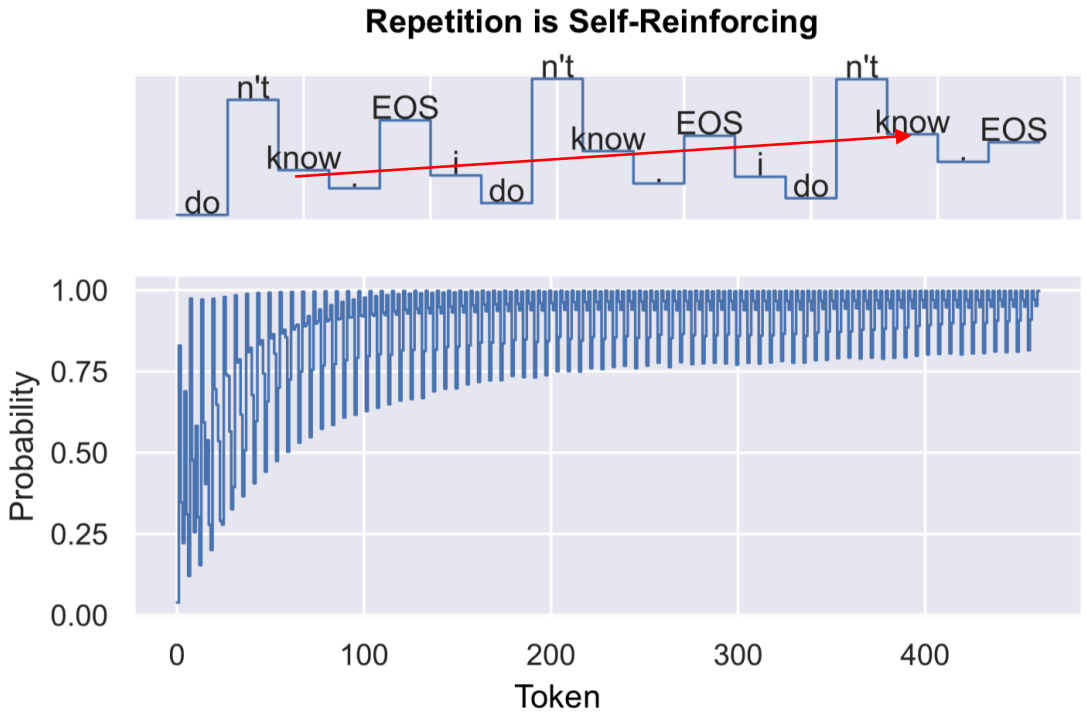
…get your hopes up. I saw him once and I have no intention of being near him anytime soon. He sat on the edge, the wind tossing around his hair. It was going to be seriously wind-blown later. I sat down next to him and I was trying to forget the dwarfs mangled body. I shook and hugged myself. Are you cold? He asked, his voice full of concern. I just shrugged and squeezed my eyes shut. I saw Kojas glowing eyes and sword, the…

BeamSearch

…looked at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds. He looks at the clouds…

The Gravitational Force of Repetition

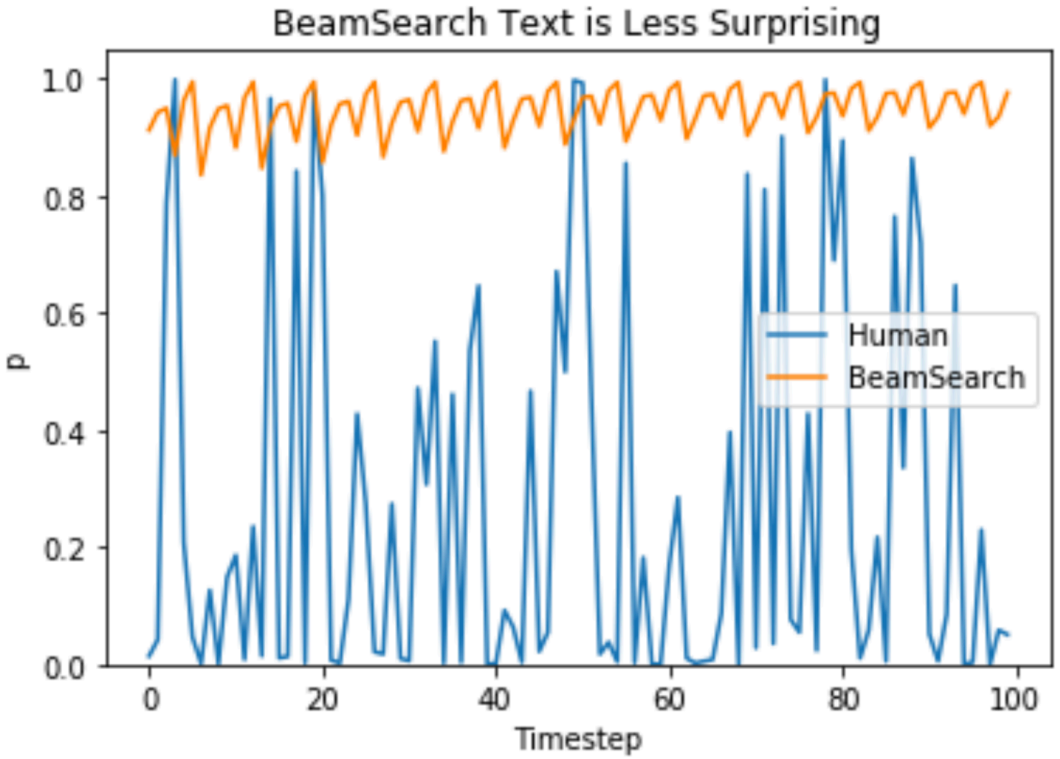
- Likelihood maximization approaches, such as **beam search**, tend to **loop** into repeating the same sentence, often a generic sentence such as *"I don't know."*



$$P(\text{“know”}|\text{“I don't”})$$
$$< P(\text{“know”}|\text{“I don't know. I don't”})$$
$$< P(\text{“know”}|\text{“I don't know. I don't know. I don't”})$$

positive-feedback loop

Natural Language Distribution has Spikes



Human

…get your hopes up. I saw him once and I have no intention of being near him anytime soon. He sat on the edge, the wind tossing around his hair. It was going to be seriously wind-blown later. I sat down next to him and I was trying to forget the dwarfs mangled body. I shook and hugged myself. Are you cold? He asked, his voice full of concern. I just shrugged and squeezed my eyes shut. I saw Kojas glowing eyes and sword, the…

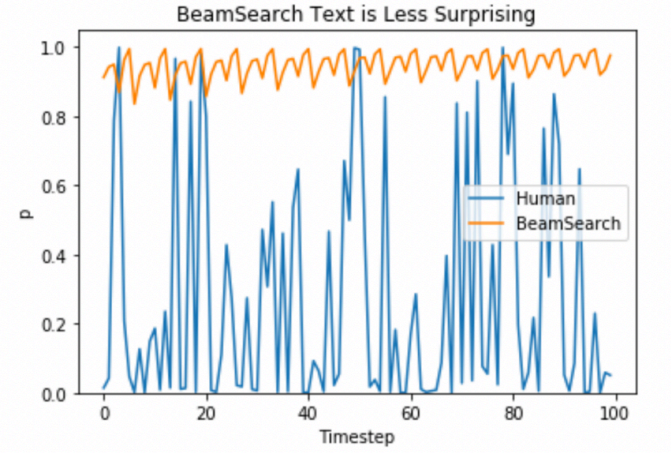
BeamSearch

…looked at the clouds. He looks at the clouds…

Figure 2: The probability assigned to tokens generated by humans and beam search using GPT-2-117M. Note the increased variance that characterizes the richness of human text.

The Turbulent Distribution of Natural Language

- Natural language rarely **remains in the high probability zone for long**, instead **dipping into the low probability zone** to give detail with content words.
- Decoding based on ***maximization*** leads to text with **unnaturally high probability** and **too little variance**.
- This motivates the use of ***randomization*** over ***maximization***, which allows us to *sample* from the model's approximation of the data distribution rather than to *optimize* output probability.



Questions

Why does decoding with **beam search** from a strong language model lead to such **degenerate** text?

Why does sampling from a truncated vocabulary distribution perform better than sampling from the whole distribution?

What is the most principled method of truncation currently available?

Why Does Sampling from the Full Distribution Lead to Degenerate Text?

🗣 **Context**: On days when he woke early, the president liked to have coffee in the oval office. There was something about watching the sky grow lighter and lighter as you sat your pajama'd behind in the most powerful chair in the free world, sipping marine strength coffee, that you just couldn't achieve anywhere else.

🥄 **Sampling** (*t*=1.0): You couldn't be sure if that's what you were really doing, and If you decided to take the day off. The president wanted you to take the day off, but he wanted to maintain a curfew and use his influence wisely.

---

incoherent

Tail of the distribution

- **“tail”** to describe the large majority of tokens, which are assigned probability that is within some small ε of 0

- **One bad sample can start a downward spiral**
  - *recency bias* and *explanation-away* problem,
  - language models have the tendency to **rely overly on the short-term context** that can easily explain away the longer-term context
- **Sampling from the tail is extremely likely**
  - in the full distribution the average probability mass assigned to the tail about 0.31.

Truncating the Distribution

- The probability of sampling from the tail goes up the more tokens are retained.

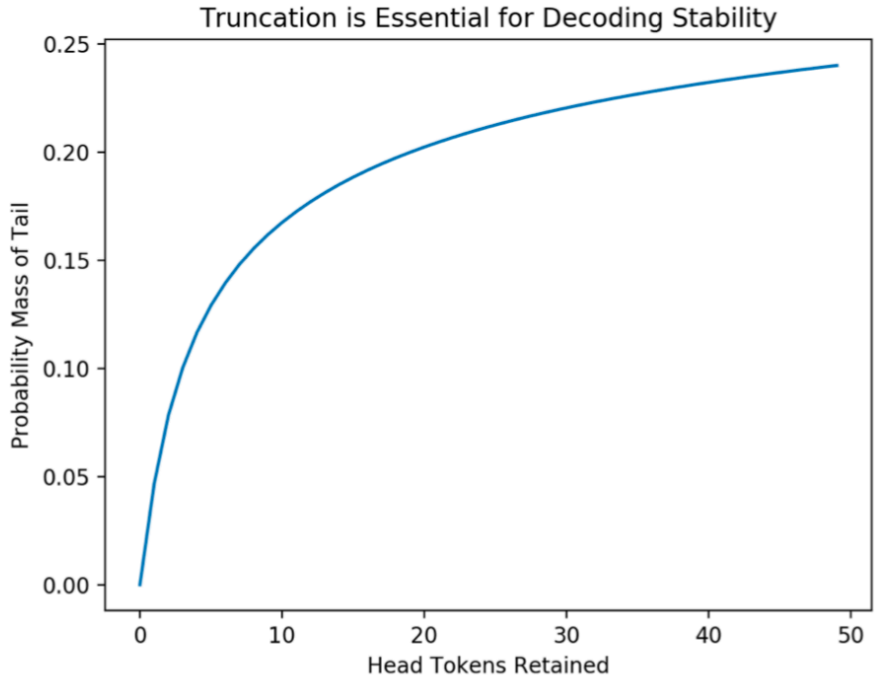


Figure 4: The chart shows the probability mass in the tail (approximated as the sum of all candidates with lower probability than the ground truth token) when only highest probability x tokens are considered; this is equivalent to asking how much of the tail is "left" when using top-k sampling where $k = x$.

Truncating the Distribution

1. Sampling with Temperature
2. Top-k Sampling

Sampling with Temperature

$$p(x = V_l | x_{1:i-1}) = \frac{\exp(u_l/t)}{\sum_{l'} \exp(u_l/t)}.$$

$t \to \infty$

greedy decoding

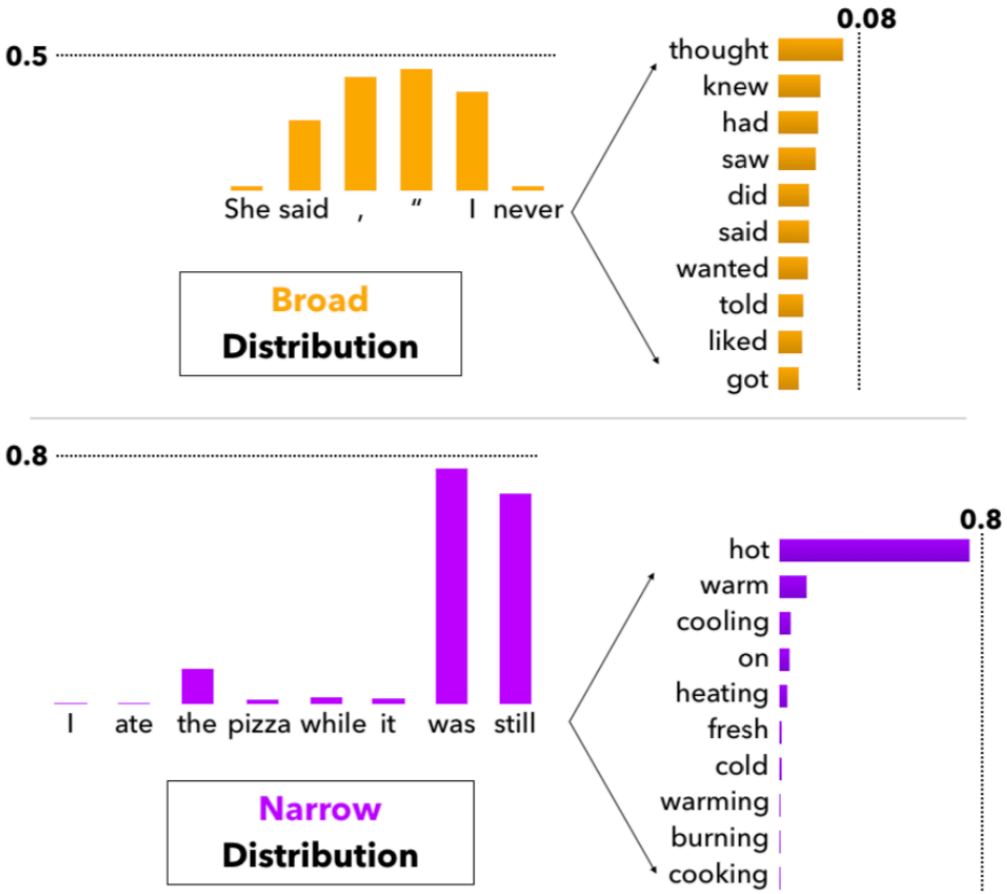
$t \to 0$

uniform sampling

Top-k Sampling

$$P'(x|x_{1:i-1}) = \begin{cases} P(x|x_{1:i-1})/p' & \text{if } x \in V^{(k)} \\ 0 & \text{otherwise} \end{cases}$$

Top-k Sampling



- flat across hundreds of reasonable options
- there are many more than k reasonable candidates, and limiting sampling to only the top-k choices runs the risk of generating bland and potentially repetitive text.

- a model may not have k reasonable candidates because the probability mass is peaked for less than k words.

Questions

Why does decoding with **beam search** from a strong language model lead to such **degenerate** text?

Why does sampling from a truncated vocabulary distribution perform better than sampling from the whole distribution?

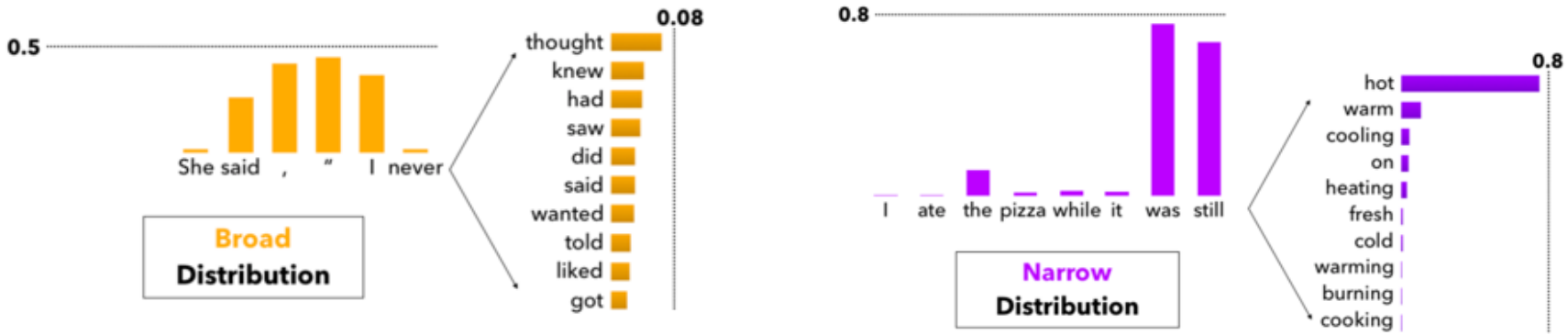
What is the most principled method of truncation currently available?

Nucleus (Top-p) Sampling

- top-p vocabulary V (p) ⊂ V is the smallest set such that：

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) >= p.$$

- In practice this means that we select the highest probability tokens whose **cumulative probability mass exceeds our pre-chosen threshold p**.



Comparison to Other Methods

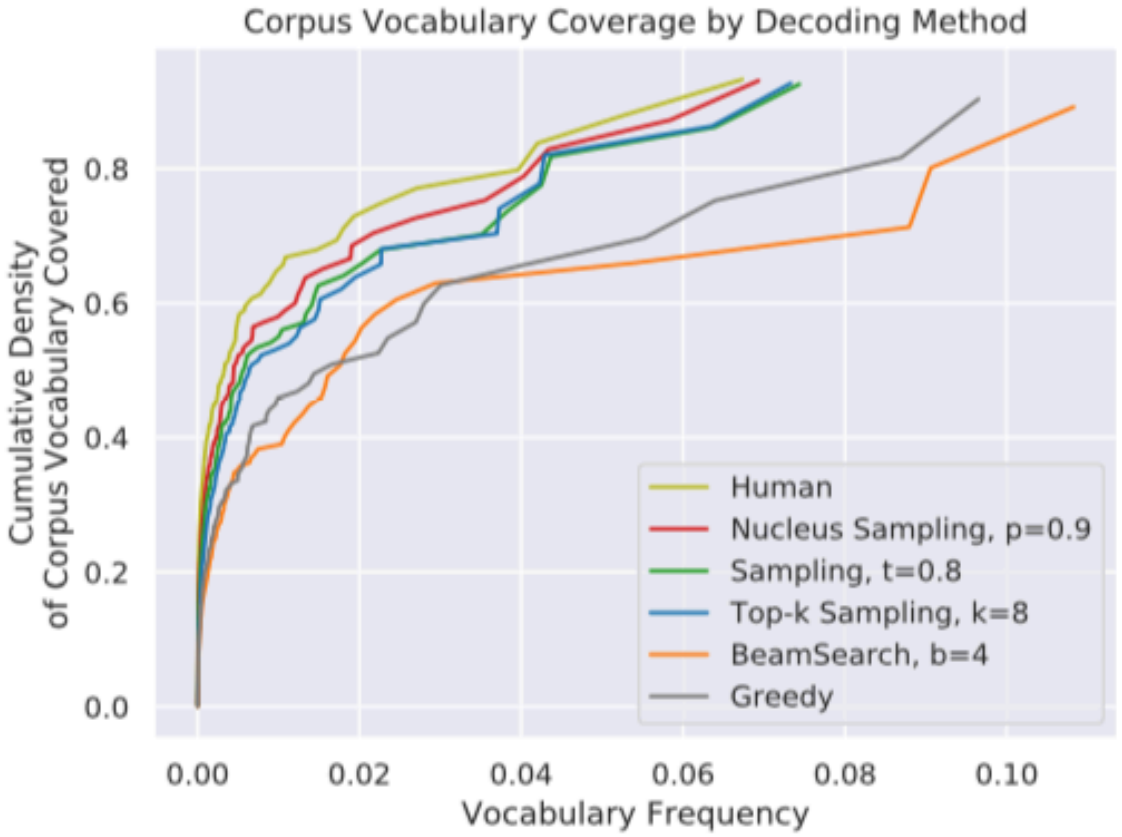
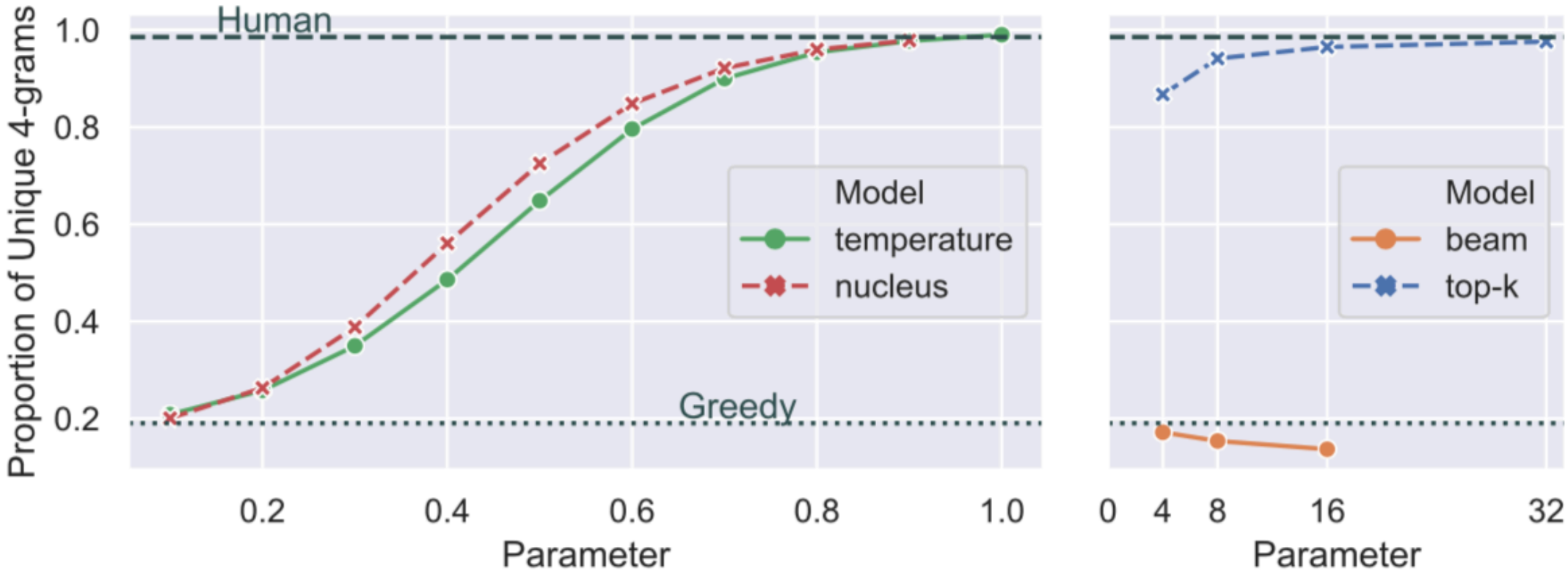


Figure 8: A chart describing the distributional differences between n-gram frequencies of human and machine text. The complete separation of likelihood maximization and stochastic methods, stochastic clearly closer to human, indicates an inherent issue with a likelihood maximization as a decoding objective.

Comparison to Other Methods



Conclusion

- Using likelihood as a *decoding objective* leads to text that is bland and strangely repetitive.
- Surprising distributional differences between human text and machine text.
- Decoding strategies alone can dramatically effect the quality of machine text, even when generated from exactly the same neural language model.
- By sampling text from the dynamic *nucleus* of the probability distribution, which allows for diversity while effectively truncating the less reliable tail of the distribution.

Thanks !