Decoding Strategies in Neural Text Generation

Xin Li

NLP Center, Tencent AI Lab

May 9, 2019

The Curious Case of Neural Text Degeneration

- Background
- Likelihood Maximization Lead to Degenerate Text
- Randomization Lead to Degenerate Text
- Advanced Strategies and Nucleus Sampling

Background

The categories of the text generation task:

- Open-ended Generation:
  - Generate text that forms a coherent continuation from the given context (A lot of plausible solutions).
  - Example applications: conditional story generation and contextual text continuation.
- Non-open-ended Generation:
  - Generate text such that the output is a close (text2text/data2text) transformation of the input (The degree of freedom is substantially less than the former).
  - Example applications: machine translation, data-to-text generation and abstractive summarization.
- Open-ended Generation and Non-open-ended Generation are not a strict dichotomy.

* In this paper, the authors mainly discuss the **Open-ended Generation** with GPT-2 (117M).

Likelihood Maximization Lead to Degenerate Text

`Beam Search` is the most popular likelihood maximization-based decoding strategies:

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…

Figure 1: Repetition Issue

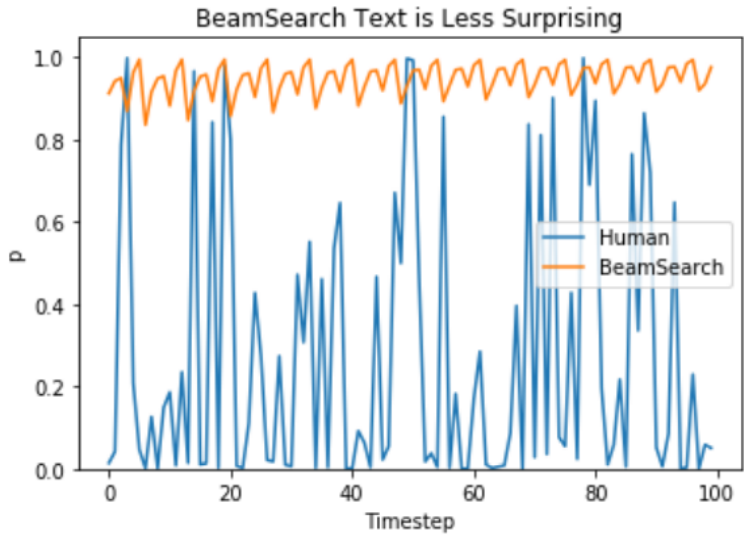


Figure 2: Striking difference between human text and machine text

- `Beam search` is still not enough.
- The differences in Fig 2 proves that the decoding strategies can dramatically affect the quality of generation even with the same neural language model.

Likelihood Maximization Lead to Degenerate Text

1. The Gravitational Force of Repetition.
   - The likelihood maximization approaches tend to loop into repeating the same sentence (often a generic sentence).
   - $P(\text{“know”} \mid \text{“I don't”}) < P(\text{“know”} \mid \text{“I don't know. I don't”}) < P(\text{“know”} \mid \text{“I don't know. I don't know; I don't ”})$
2. The Turbulent Distribution of Natural Language.
   - Human rarely remains in high probability zone for long time.
   - The maximization-based methods force the model to obtain cumulative likelihood as high as possible.

Human	BeamSearch
…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…	…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…

3. In summary, maximization decoding leads to text with unnaturally high probability and too little variance.

Randomization instead of Maximization

- Idea: *Sampling* from the model's approximation of the data distribution rather than *maximizing* output probability.
- However, sampling from the full distribution still lead to degenerate text.
  - i) One bad sampling can start a incoherent generation.
  - ii) Sampling from the tail is extremely likely:

$$P(\bar{A}) = \prod_{i=n+1}^{n+m} (1 - \epsilon_i)$$

🗣 **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.

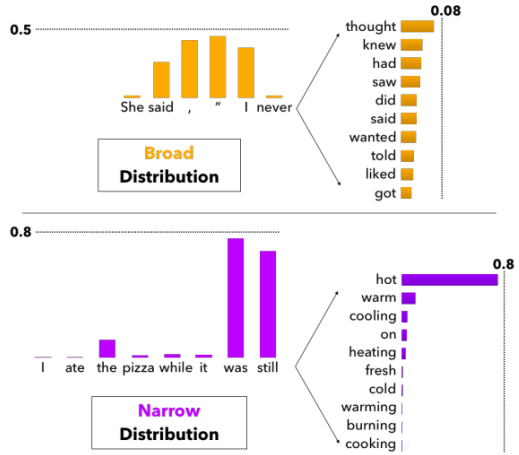
Sampling with Truncation

The simplest solution to dealing with the tail words is to only focus on the most important tokens.

- Sampling with Temperature (soft): reshape the distribution with constant $t \in [0, 1)$, skewing towards high probability events.
- Top-k Sampling: sampling from the k tokens with highest probability mass.

Top-k sampling recently become popular but it still sub-optimal

- In a flat distribution, there are many more than k candidates
- In the opposite scenario, the probability mas is peaked for less than k words.



Sampling with Truncation

Dynamically adjusting the value of k should be ok:

- Based on this idea, the author propose Nucleus Sampling (Top-p Sampling).
  - Select the minimal subset of the vocabulary whose cumulative probability mass exceeds the pre-chosen threshold p.
  - Sampling from the selected subset.

Comparison

🗣️ **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.

💰 **Greedy**: He'd been in the oval office for over a year now, and he'd never been so happy to be there. He'd been in the oval office for almost a year, and he'd never been so happy to be there.

🚦 **BeamSearch** (*k*=4): The President was a man of his word. He was a man of his word. He was a man of his word. He was a man of his word. He was a man of his word.

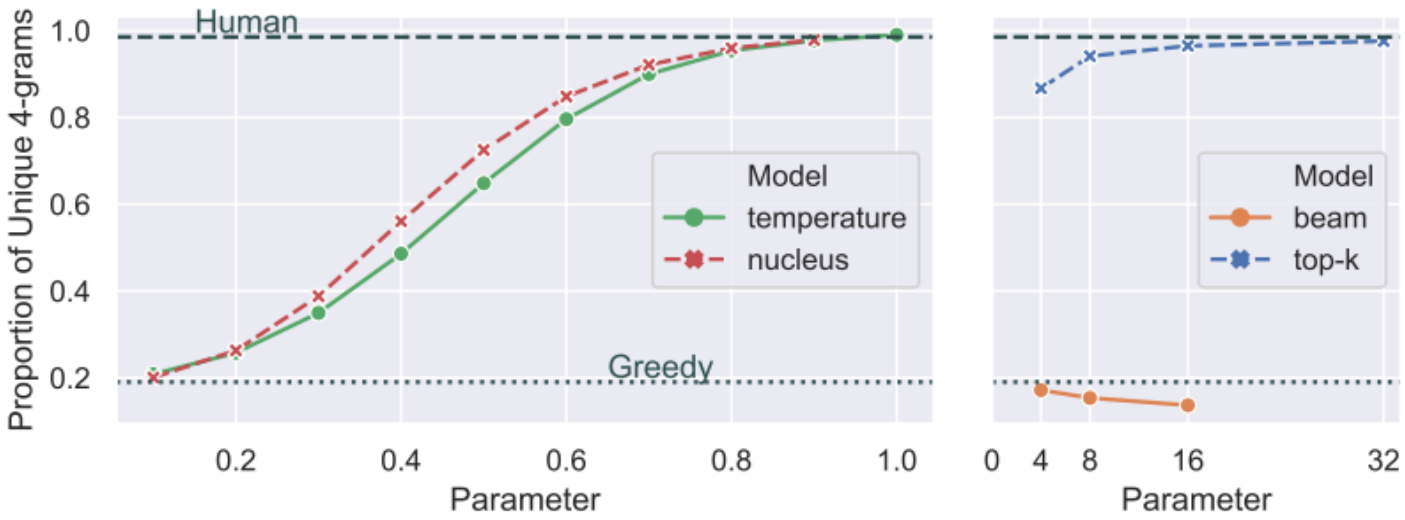
🥄 **Sampling** (*t*=1.0): You couldn't be sure if that's what you were really doing, and If you decided to take the day off. The president wanted you to take the day off, but he wanted to maintain a curfew and use his influence wisely.

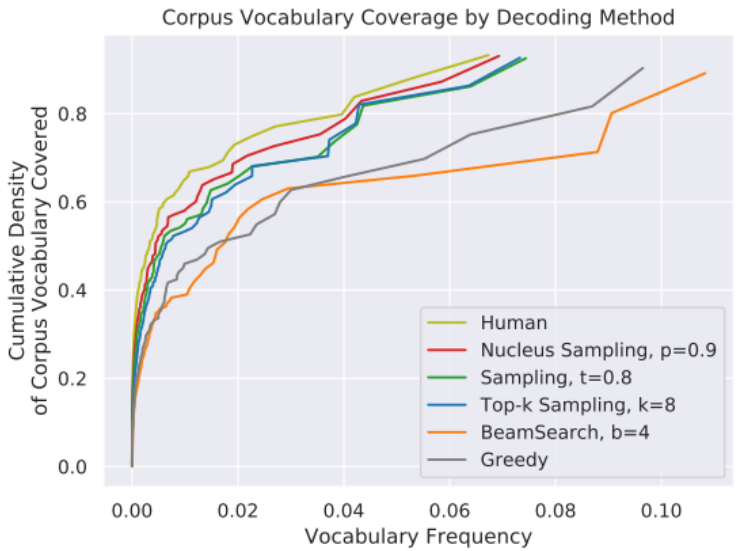
🇰 **Top-*k* Sampling** (*k*=8): "I'm sorry, sir." "No, it's okay, son," the president said. "I understand." "You're going to have to make a special trip down there to get that kid. He has no idea where he's going."

⚛️ **Nucleus Sampling** (*p*=0.9): But that wasn't what drew the president's attention. He'd been seated for maybe a minute when he noticed the other man. What was the guy doing here?

🗣️ **Gold**: He was therefore disagreeably surprised to find a man in an understated grey suit sitting in that selfsame chair sipping tea. The president turned around and went looking for his chief of staff.

Comparison


Human
Greedy
Proportion of Unique 4-grams
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
Parameter
Model
temperature
nucleus
0
4
8
16
32
Parameter
Model
beam
top-k


Corpus Vocabulary Coverage by Decoding Method
Cumulative Density
of Corpus Vocabulary Covered
0.0
0.2
0.4
0.6
0.8
0.00
0.02
0.04
0.06
0.08
0.10
Vocabulary Frequency
Human
Nucleus Sampling, p=0.9
Sampling, t=0.8
Top-k Sampling, k=8
BeamSearch, b=4
Greedy