

THE CURIOUS CASE OF
NEURAL TEXT
*De*GENERATION

Морозов Дмитрий, БПМИ172

Наиболее известные методы декодирования

- Likelihood Maximization (Beam Search)
- Top-k
- Pure Sampling
- Sampling with Temperature
- ...
- Nucleus Sampling?

Проблемы существующих методов декодирования

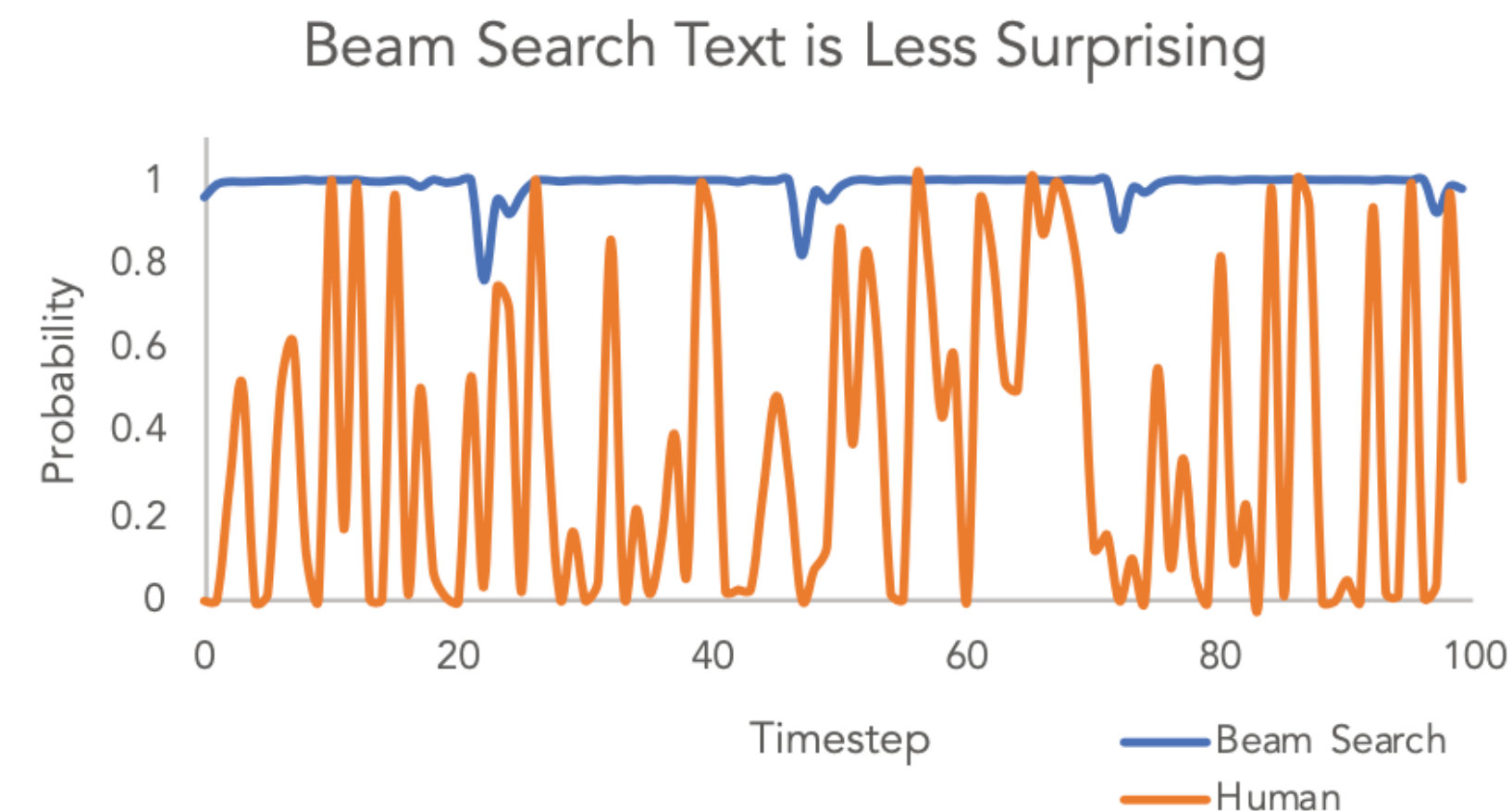
Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, $b=32$:

[illegible]

Pure Sampling:

They were cattle called **Bolivian Cavalleros**; they live in a remote desert **uninterrupted by town**, and they speak **huge, beautiful, paradisiacal Bolivian linguistic thing**. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. **"They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."**



Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Open-Ended VS Directed Generation

Open-Ended generation:

$$P(x_{1:m+n}) = \prod_{i=1}^{m+n} P(x_i | x_1 \dots x_{i-1})$$



WebText



Beam Search, $b=16$



Pure Sampling



Sampling, $t=0.9$



Top-k, $k=640$



Top-k, $k=40$, $t=0.7$



Nucleus, $p=0.95$



WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on [the West Australian coast increasing by more than 50 per cent in the past year](#). The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

The Australian Food Safety Authority has warned Australia's beaches may be **revitalised** this year because healthy **seabirds and seals** have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by **the Holden CS118 and Adelaide Airport CS300 from 2013**. A major **white-bat and umidauda** migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: **packed in the belly of one killer whale thrashing madly** in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, **he'd been seen tagged for a decade**.

Pumping Station #3 shut down due to construction damage Find more at:

www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html

"In the top 10 killer whale catastrophes in history:

1) 1986: Up to 12 orcas **struck by lightning; many drowned** and many more badly injured.

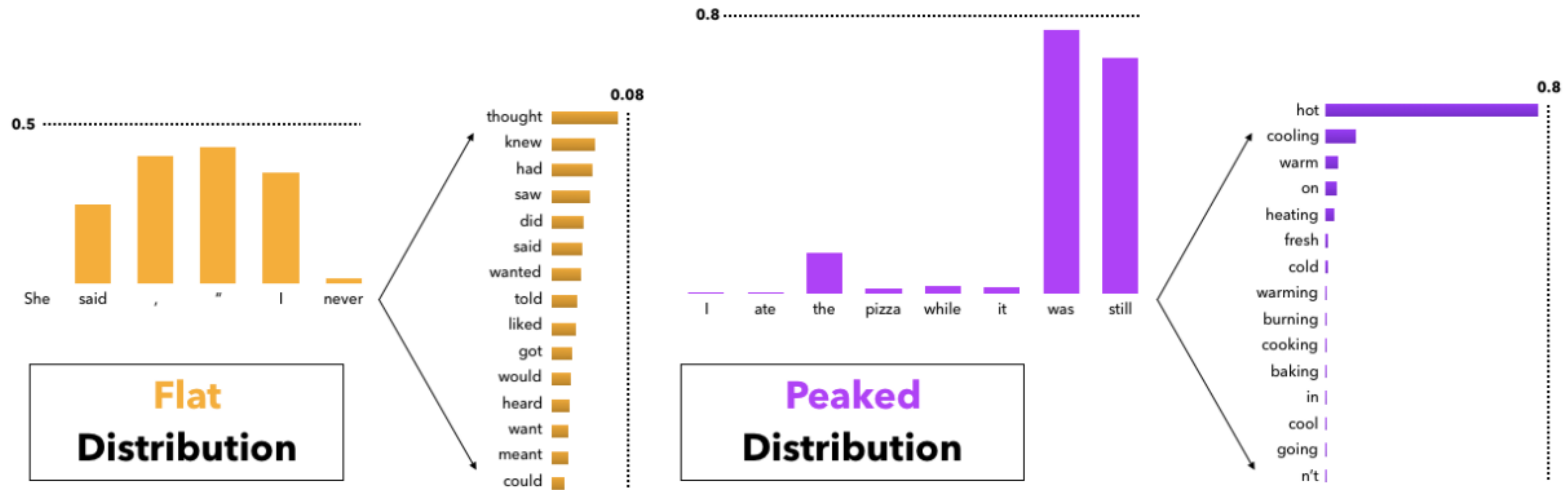
The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a **fishing vessel off the coast of Bundaberg**, and died after being **sucked into the ocean**. **The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.**

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the **Petrels** are shrinking and dwindling population means there will only be room for a few **new fowl**.

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Описание методов декодирования

Top-k:



Sampling with Temperature:

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

Описание методов декодирования

Nucleus Sampling:

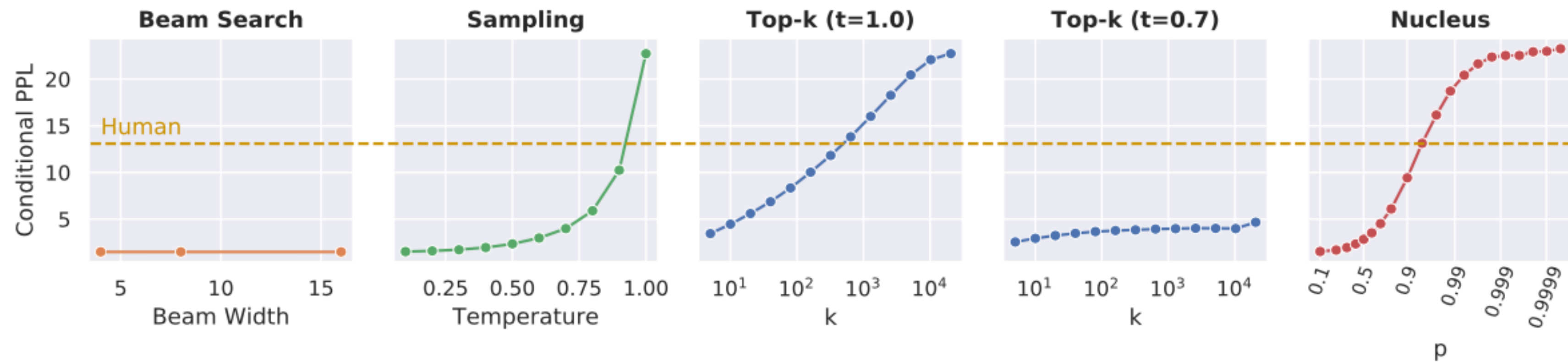
$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p.$$

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

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

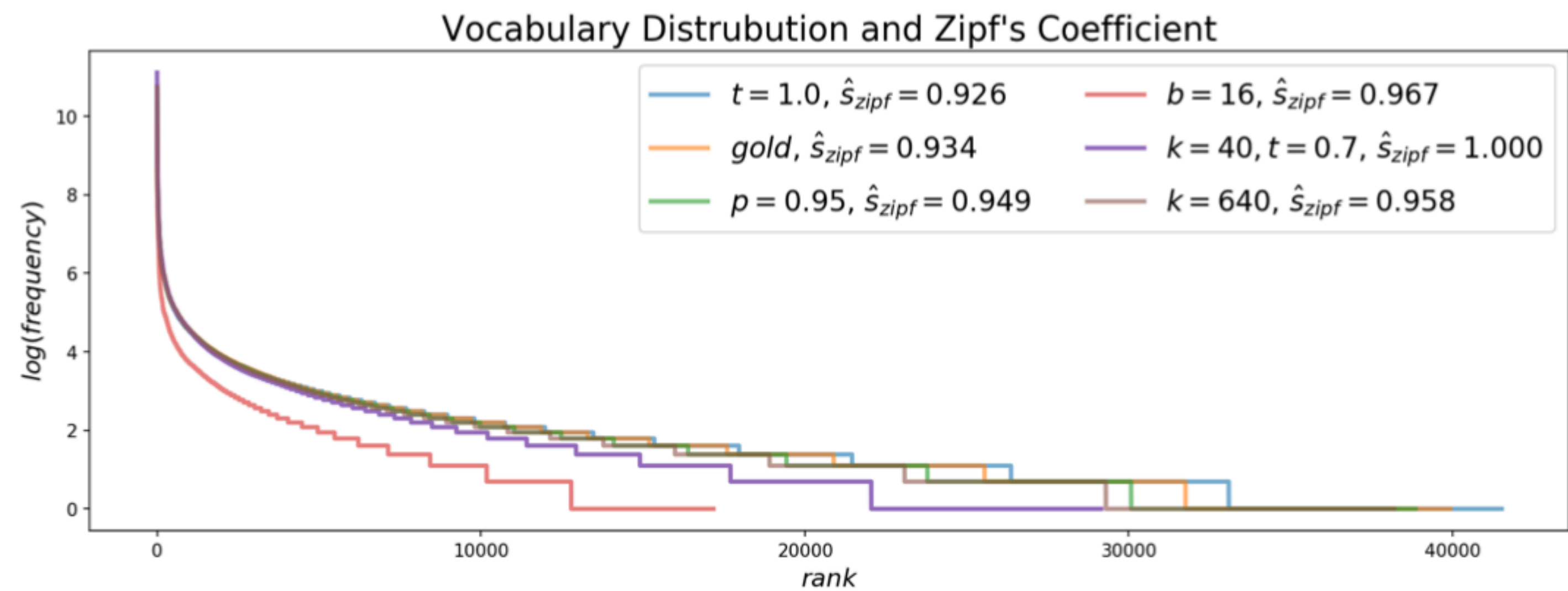
Сравнение методов декодирования

Сложность сгенерированного текста (perplexity)



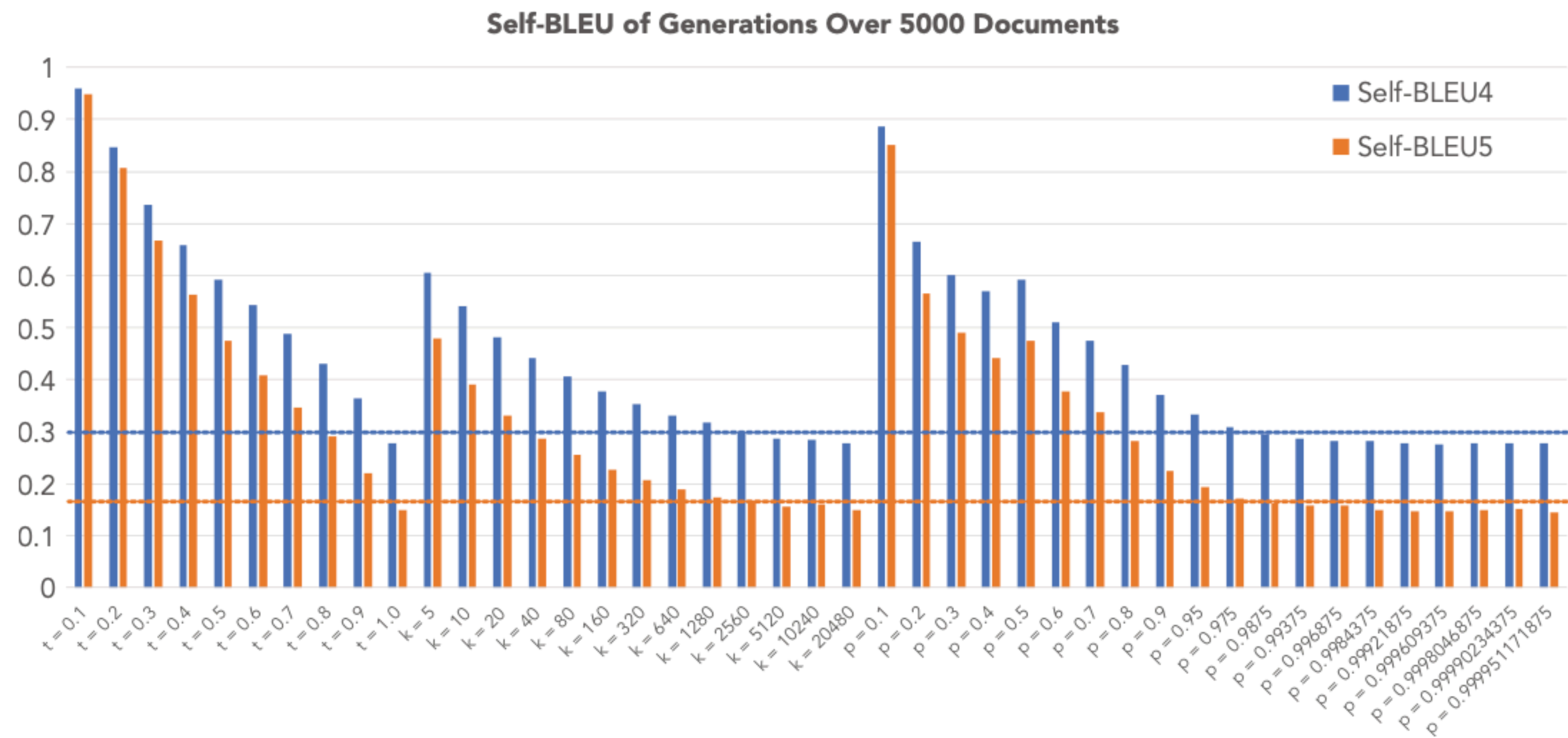
Статистическая оценка распределения

Анализ распределения Zipf (словарный запас)



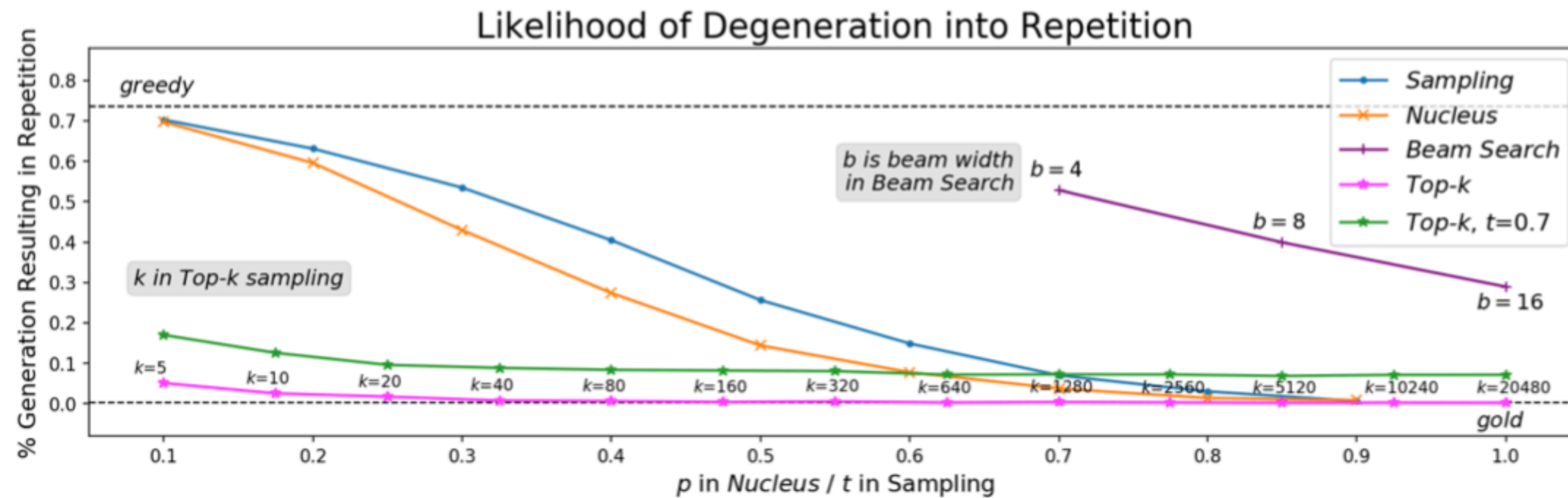
Статистическая оценка распределения

SELF-BLEU (разнообразие)



Статистическая оценка распределения

Повторение слов



Статистическая оценка распределения

Human Unified With Statistical Evaluation (HUSE)

Method	Perplexity	Self-BLEU4	Zipf Coefficient	Repetition %	HUSE
Human	12.38	0.31	0.93	0.28	-
Greedy	1.50	0.50	1.00	73.66	-
Beam, b=16	1.48	0.44	0.94	28.94	-
Stochastic Beam, b=16	19.20	0.28	0.91	0.32	-
Pure Sampling	22.73	0.28	0.93	0.22	0.67
Sampling, $t=0.9$	10.25	0.35	0.96	0.66	0.79
Top- $k=40$	6.88	0.39	0.96	0.78	0.19
Top- $k=640$	13.82	0.32	0.96	0.28	0.94
Top- $k=40$, $t=0.7$	3.48	0.44	1.00	8.86	0.08
Nucleus $p=0.95$	13.13	0.32	0.95	0.36	0.97

СПИСОК ИСТОЧНИКОВ

- <https://arxiv.org/abs/1904.09751>

Список вопросов

1. Какие проблемы наиболее характерны для методов на основе максимизации правдоподобия (например, поиск лучом), а какие для чистой выборки (Pure Sampling)?
2. В чем заключается идея подхода Nucleus Sampling? Запишите формулу вероятности для его предлагаемого распределения.
3. В чем заключается идея применения анализа распределения Zipf и что такое Self-BLEU?