The Curious Case Of Neural Text DeGeneration

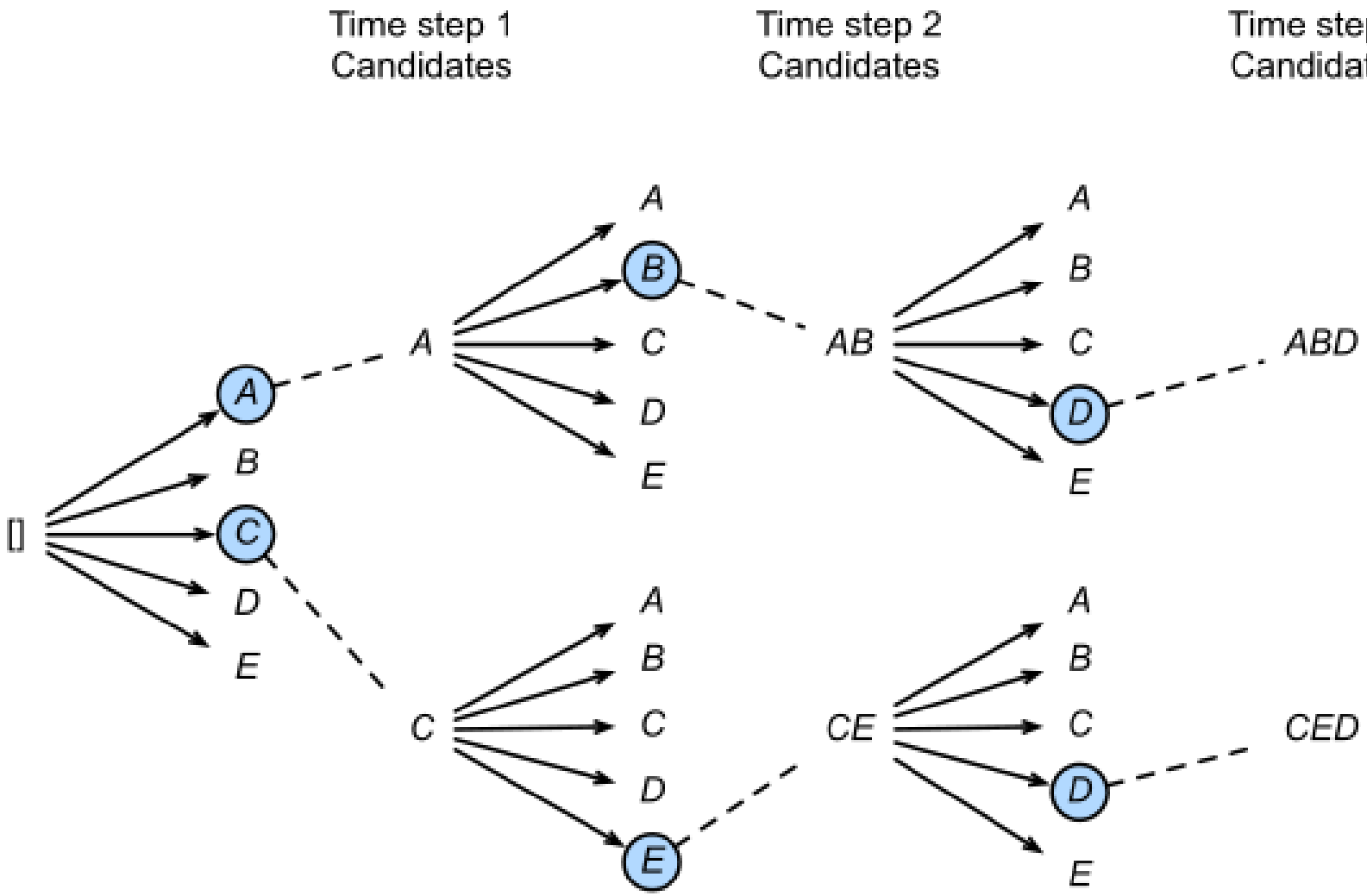
Шабалин Александр

18 ноября 2020 г.

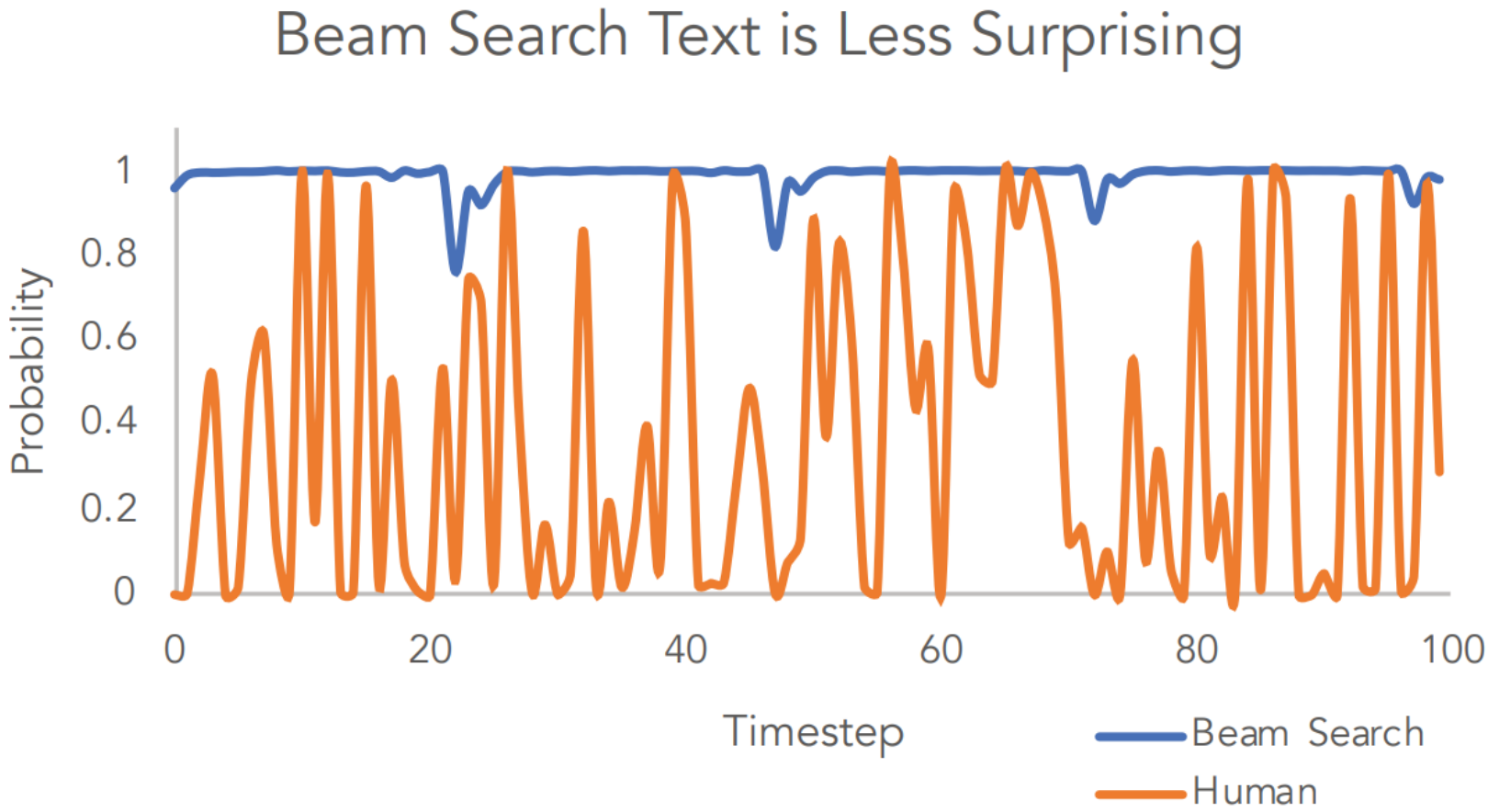
Введение

- Решаем задачу генерации текста
- Обучили языковую модель
- Хотим получать слова из распределения, чтобы текст получался похожим на человеческий.

Beam search

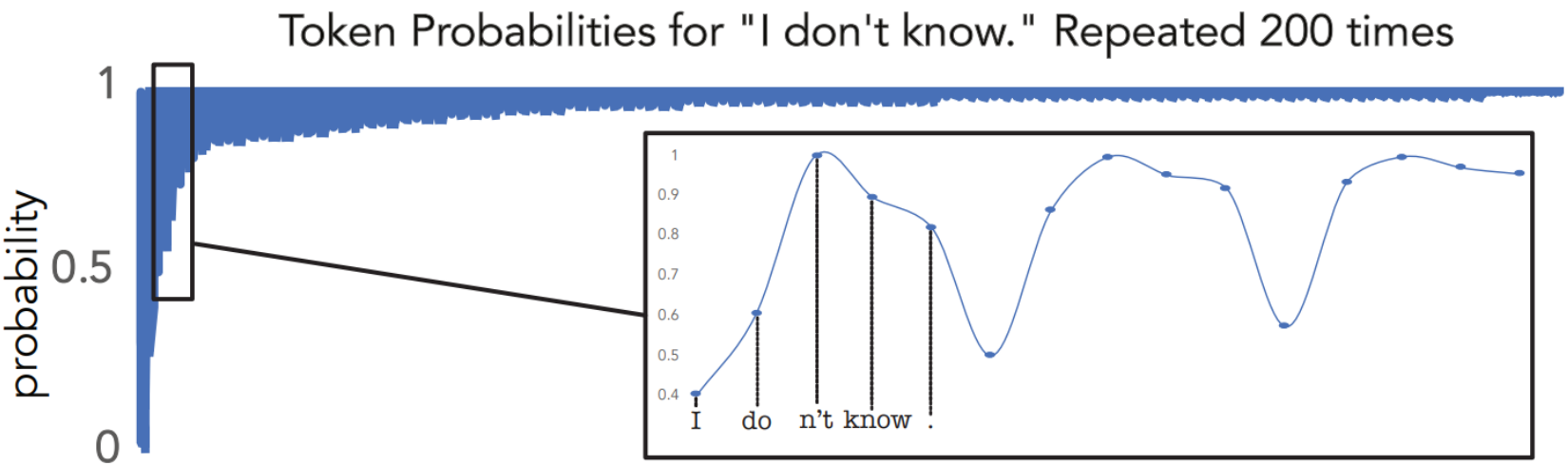

Time step 1
Candidates
Time step 2
Candidates
Time step 3
Candidates
[]
A
B
C
D
E
A
B
C
D
E
AB
ABD
C
CE
CED

Beam search



Максимизация правдоподобия не подходит для генерации текста.

Beam search



Повторяемый текст более вероятный.

Из-за этого beam search работает плохо.

Pure sampling

Будем семплировать из распределения, не выбратывая ничего. Из-за хвостов в распределении текст становится бессвязным.

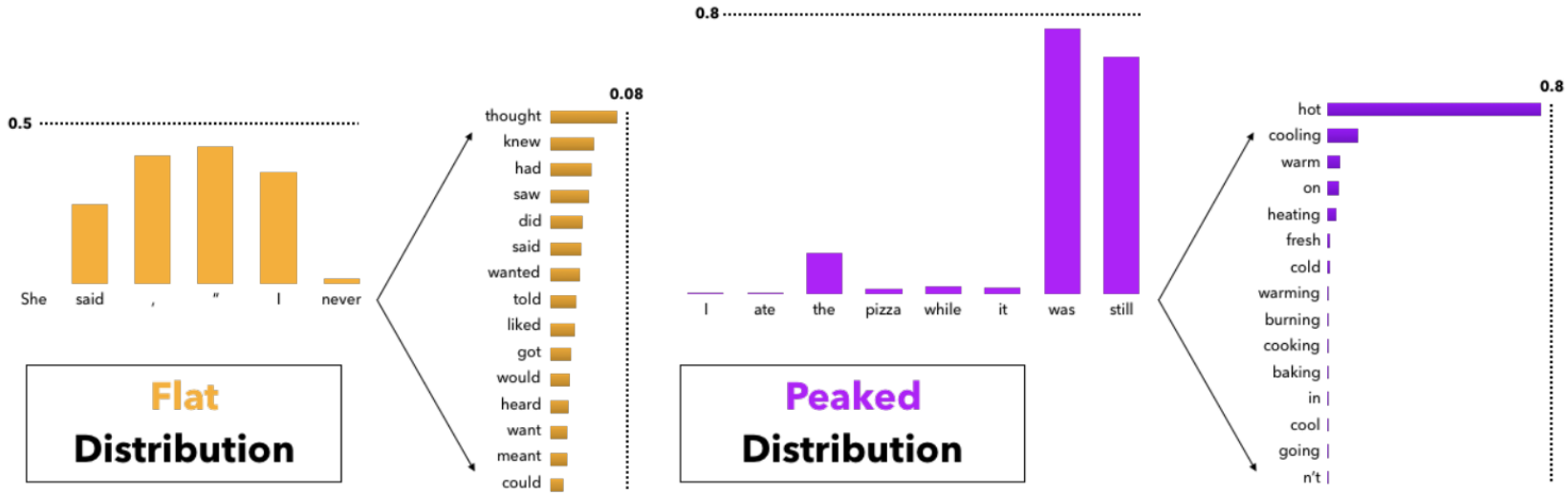
Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

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."

Необходимо выбрасывать слова с очень маленькими вероятностями.

Top-k sampling

Теперь будем брать k наиболее вероятных слов из распределения и семплировать из них.
Проблема: используем фиксированное k, которое не учитывает форму распределения.



Sampling with temperature

$$p(x = V_l) = \frac{\exp(u_l/t)}{\sum_{l'} \exp(u_{l'}/t)}$$

V — словарь, u — выход модели

При $t \in [0, 1)$ распределение становится более вырожденным.

Sampling with temperature

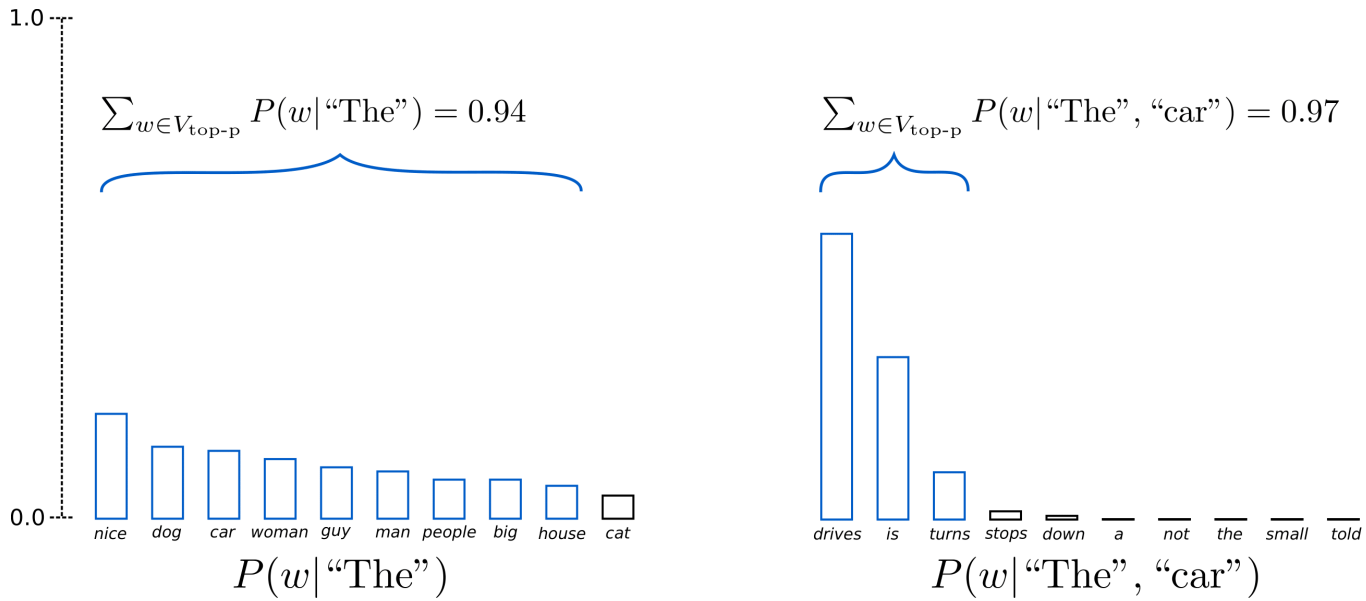
$$p(x = V_l) = \frac{\exp(u_l/t)}{\sum_{l'} \exp(u_{l'}/t)}$$

V — словарь, u — выход модели

При $t \in [0, 1)$ распределение становится более вырожденным.

Проблема: уменьшается разнообразие текста.

Nucleus (top-p) sampling



$$\sum_{x \in V^{(p)}} P(x) \geq p, \quad V^{(p)} - \text{наименьшее подмножество словаря}$$

Теперь мы учитываем форму распределения, и проблема с хвостами исчезает.

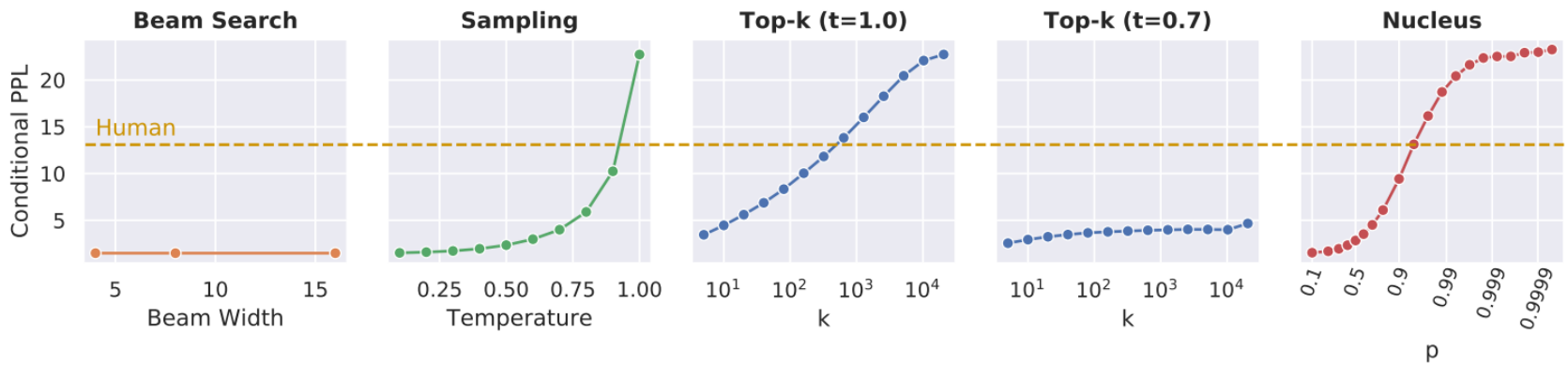
Анализ и эксперименты

Прейдем к сравнению методов.

Перплексия

$$PPL(X) = \exp\left(-\frac{1}{t}\sum_{i=1}^{t}\log p(x_i \mid x_{<i})\right)$$

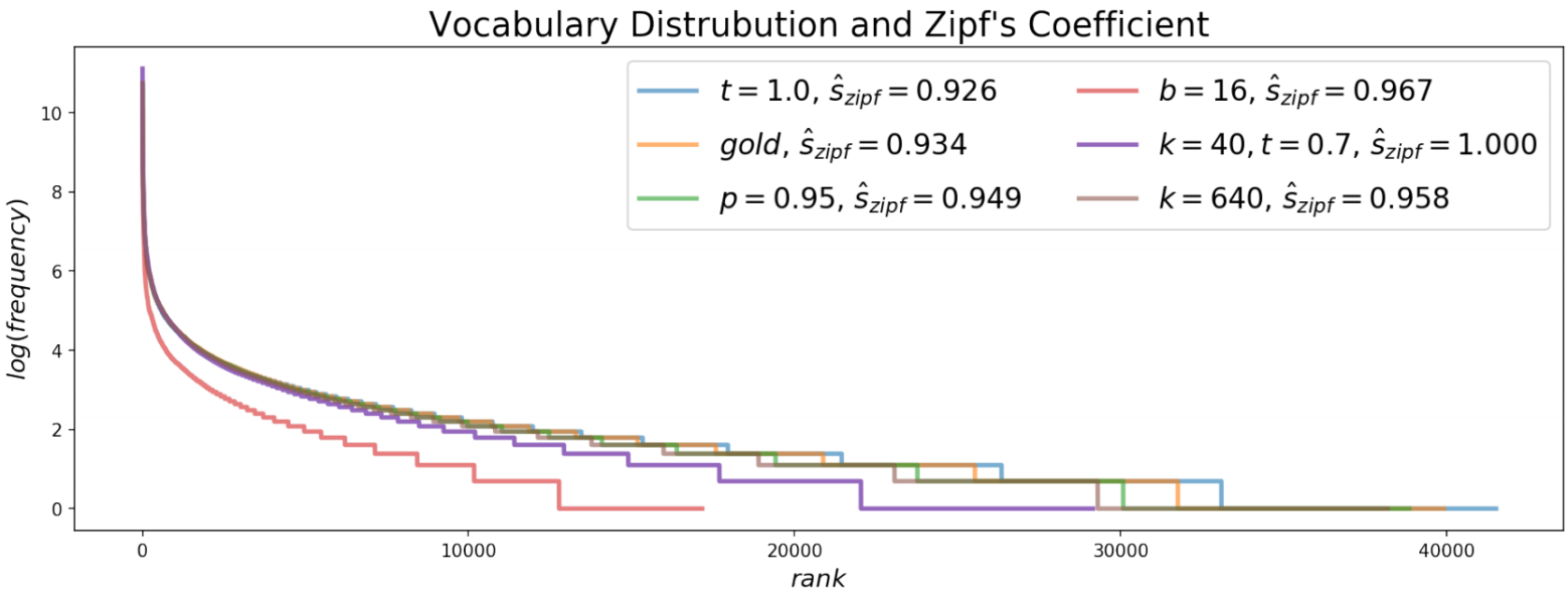
Показывает вероятность текста (чем меньше, тем вероятнее).



Лучше всего себя показывают top-k и nucleus sampling.

Распределение Ципфа

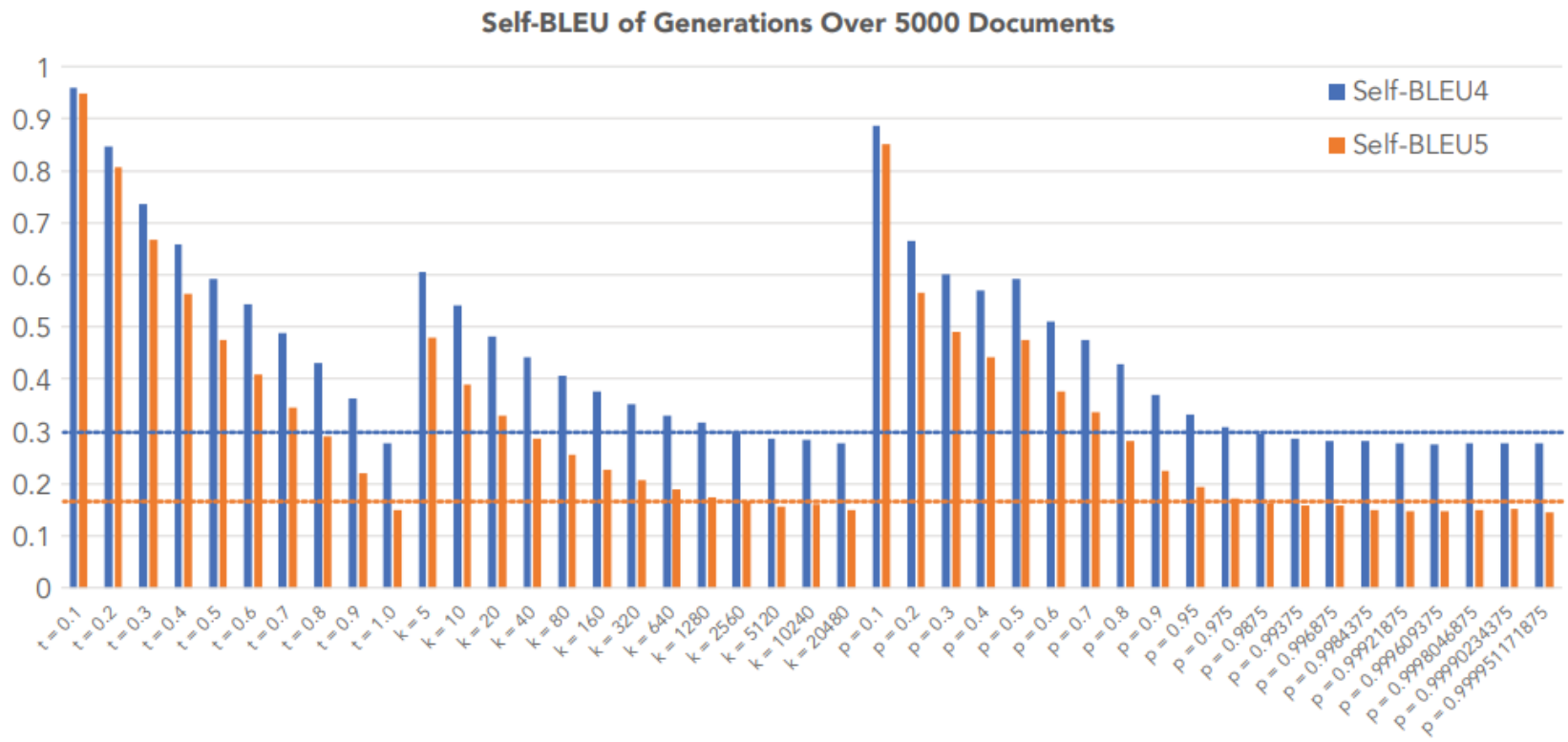
$$freq(x_{(n)}) = \frac{1}{n^s} freq(x_{(1)})$$



По частоте слов лучшим оказывается pure sampling, но top-k и nucleus тоже хороши.

Self-BLEU

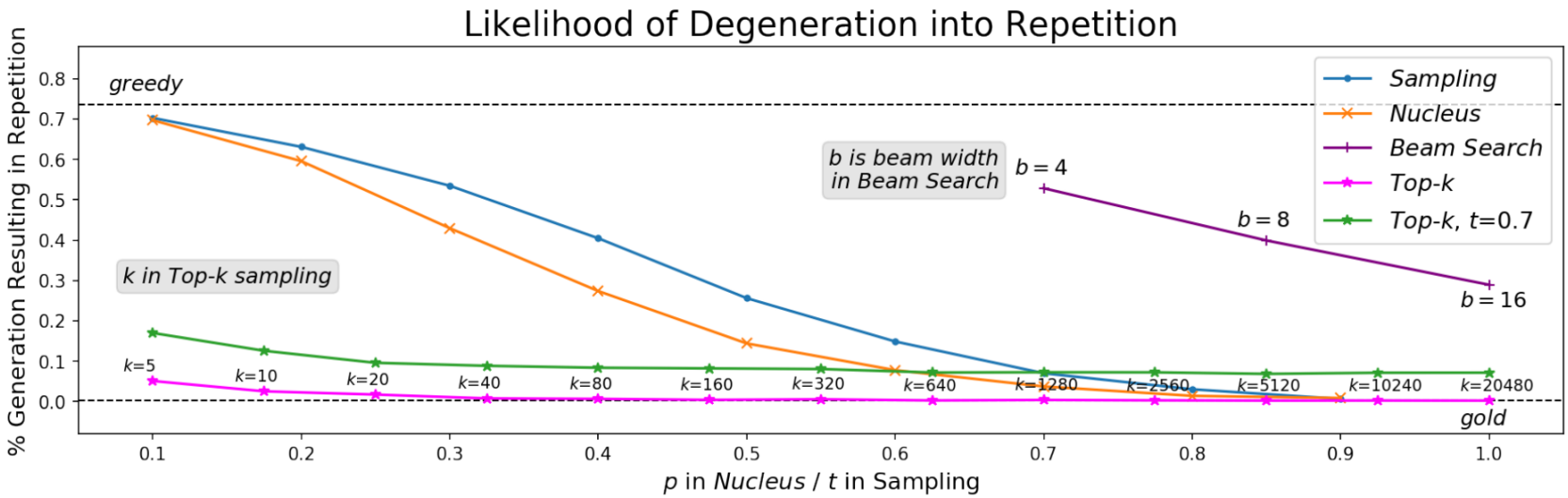
Показывает разнообразия текста.



Обычно $t \in [0.5, 1]$, $k \in [1, 100]$, а $p \in [0.9, 1)$.
Поэтому nucleus sampling лучше.

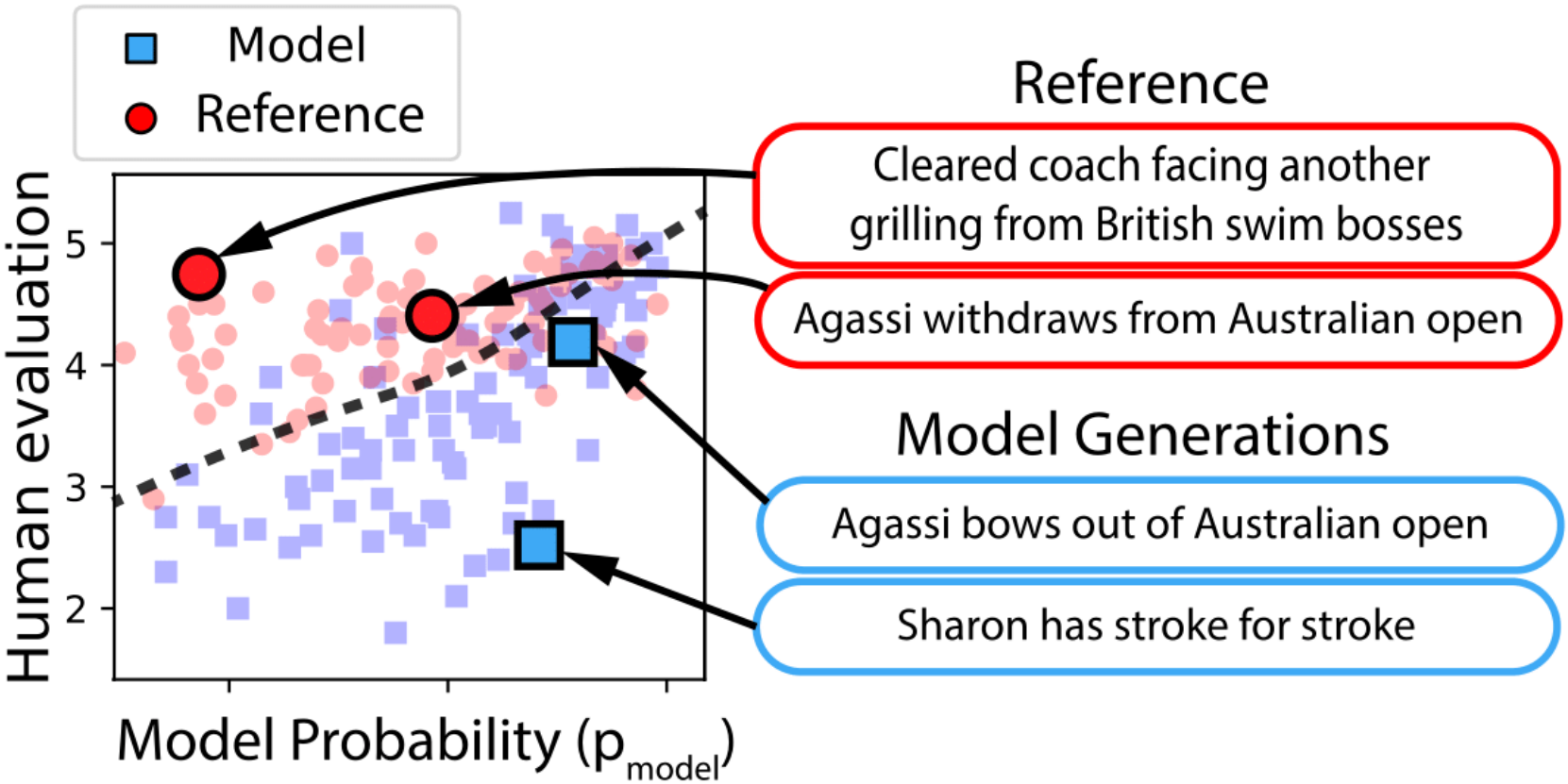
Repetition

Генерируем 200 слов.
Фраза (> 1 слова) считается повторением, если она встречается хотя бы 3 раза.



Опять побеждают те же алгоритмы.

HUSE



$$L_{HUSE} = 2 \times KNN_{error}$$

Results

Сравним методы по всем метрикам с человеческим текстом.

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**

Предложенный метод семплирования показывает себя лучше всего по совокупности метрик.

Questions

1) В чем заключается проблема вероятностного семплирования из распределения, полученного от языковой модели? Предложите варианты ее решения.

2) В чем заключается главная проблема top-k sampling, и как Nucleus Sampling ее решает?

3) Почему плохо семплировать слова, опираясь только на максимизацию правдоподобия?

Bibliography

1) The Curious Case of Neural Text Degeneration

2) HUSE