DSAIL

Data Science Artificial Intelligence Lab.

2022 Fall Lab Seminar

- The Curious Case of Neural Text DeGeneration (2019, ICLR)

2022.11.24
Jeon Hyolim
gyfla1512@g.skku.edu

Content

- Text Generation Method
- Decoding Strategy Method
- Maximization-based Method
- Sampling-based Method
- Evaluation

Text Generation Method

Directed Generation VS Open-Ended Generation

1. Directed Generation

- Text Generation: (input, output) pairs

- **Task**: machine translation (Bahdanau et al., 2015), data-to-text generation (Wiseman et al., 2017), summarization (Nallapati et al., 2016)

- **Method**: with an attention mechanism encoder-decoder architectures (Bahdanau et al., 2015; Luong et al., 2015), using attention-based architectures such as the Transformer (Vaswani et al., 2017).

Text Generation Method

Directed Generation VS Open-Ended Generation

2. Open Ended Generation

- (1) conditional story generation
  (2) contextual text continuation

- Recently become a **promising research direction** due to significant advances in neural language models (Clark et al., 2018; Holtzman et al., 2018; Fan et al., 2018; Peng et al., 2018; Radford et al., 2019)

- given a sequence of m tokens x1 . . . xm as context, the task is to generate the next n continuation tokens to obtain the completed sequence x1 . . .xm+n.

Decoding Strategy Method

- what is the best **decoding strategy** is for text generation from a language model(e.g. to generate a story)

- **Maximization based**
  Greedy Search, Beam search

- **Neural based**
  Sampling with Temperature, Top-k Sampling,
  Nucleus Sampling(Top-p Sampling)

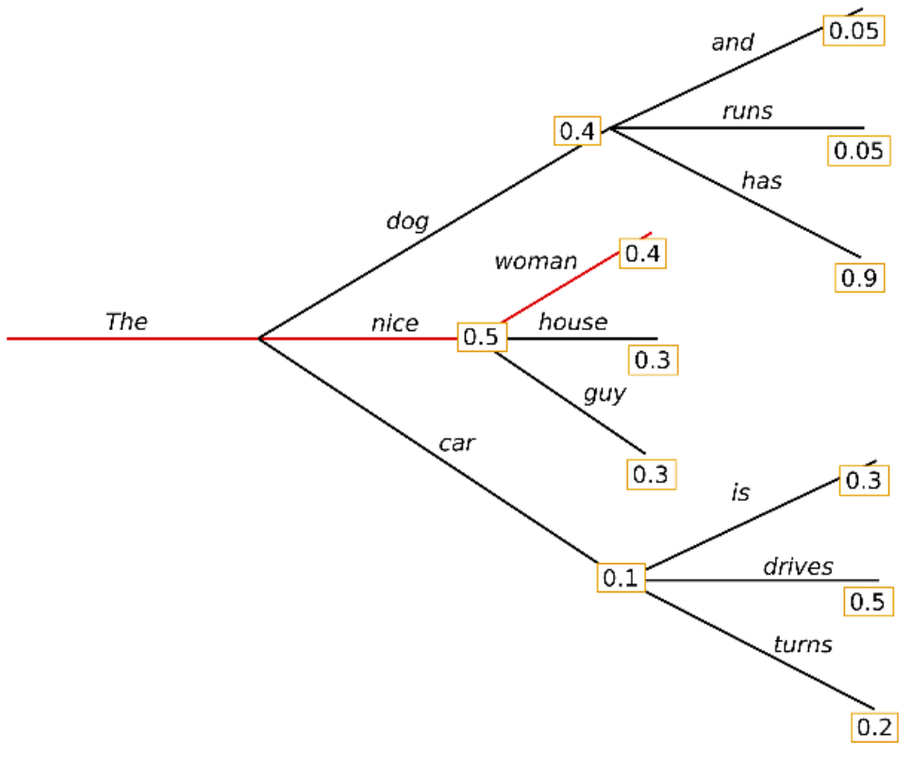
Maximization-based Method

- most commonly used decoding objective, in particular for directed generation
- Assumes that the model assigns higher probability to higher quality text, these decoding strategies search for the continuation with the highest likelihood

Maximization-based Method

1. Greedy Search

- Simple, intuitive decoding strategy
- Simply take the **highest probability** word at each position in the sequence and predict that in the output sequence.
- Time complexity(O), Accuracy(x)



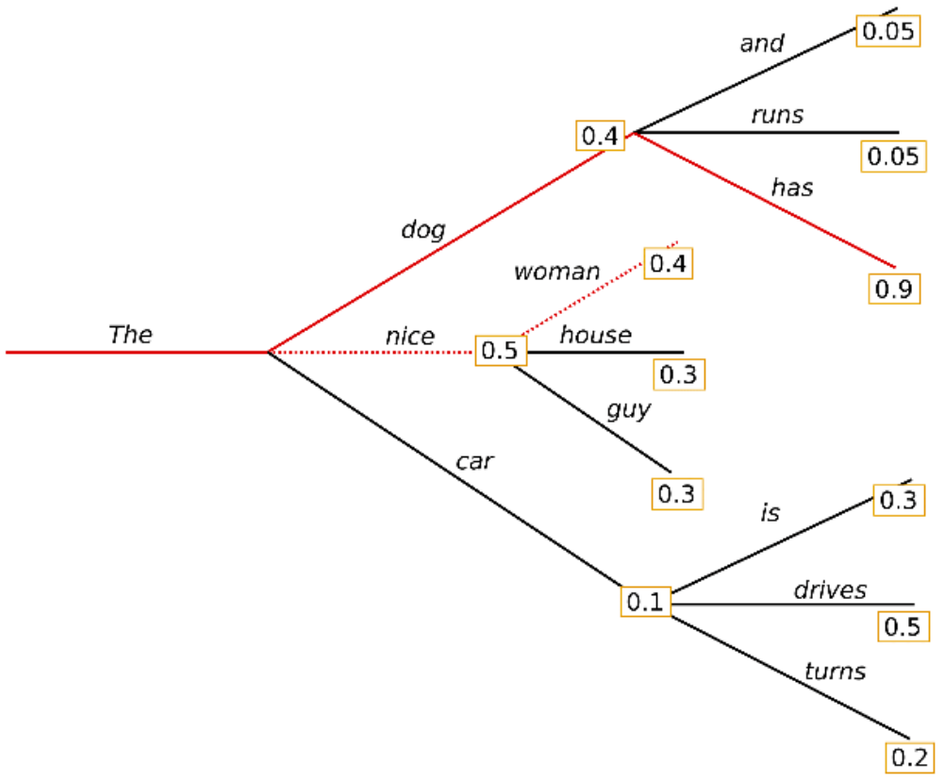

```
# encode context the generation is conditioned on
def tokenizing(text):
    return torch.tensor(tokenizer.encode(text,
                    add_special_tokens=False).ids).unsqueeze(0).to('cuda')

input_ids = tokenizing("이순신은 조선 중기의 무신이다.")

# 생성 모델은 generate 함수를 통해 다음 token을 생성 가능.
greedy_output = model.generate( input_ids, max_length=50)
# 이 안에 들어가는 값들이 decoding 옵션들
# generate text 는 context length를 포함해서 max_length에 도달하기 전까지 문장을 생성

print(tokenizer.decode(greedy_output.tolist()[0], skip_special_tokens=True))
```

```
이순신은 조선 중기의 무신이다.</s><s> 이 목록은 대한민국의 음악인  작곡가, 작사가, 작
```

Maximization-based Method

2. Beam Search

- Improved version of greedy search
- Selects multiple tokens(beam size(k), k best alternatives) for a position in a given sequence based on conditional probability




```
# activate beam search and early_stopping
beam_output = model.generate(
    input_ids,
    max_length=50,
    num_beams=5,
    early_stopping=True
)

print(tokenizer.decode(beam_output.tolist()[0], skip_special_tokens=True))
```

```
이순신은 조선 중기의 무신이다.</s><s> 그 후, 그는 《삼국사기》(三國史記)와
```

Maximization-based Method

Limitation

- incredibly **degenerate**, even when using state-of-the-art models such as GPT-2 Large

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."

Figure 1: Even with substantial human context and the powerful GPT-2 Large language model, Beam Search (size 32) leads to degenerate repetition (highlighted in blue) while pure sampling leads to incoherent gibberish (highlighted in red). When $b \geq 64$, both GPT-2 Large and XL (774M and 1542M parameters, respectively) prefer to stop generating immediately after the given context.

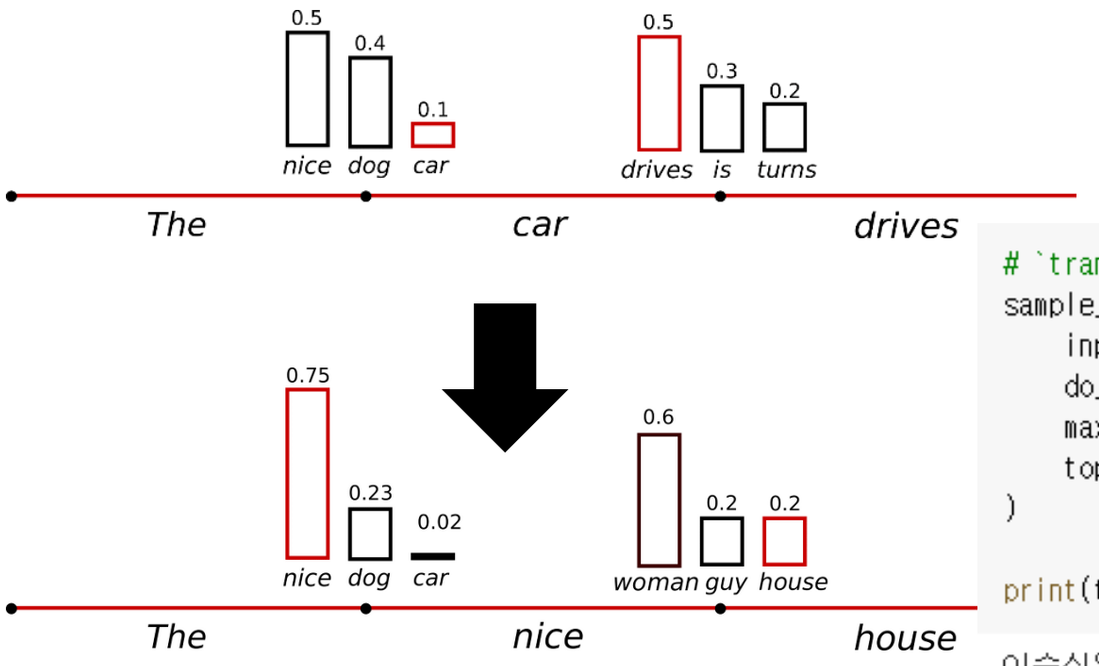
Sampling-based Method

1. Sampling

- Sampling

- sampling directly from the probabilities predicted by the model
- Probability of generating text with various words increases, but the probability of generating awkward text also increases.

- Adjust the probability distribution: "temperature“




```
# `transformers`에서 `do_sample=True`를 설정하고 `top_k=0`을 통해 *Top-K* sampling을 비활성화.
sample_output = model.generate(
    input_ids,
    do_sample=True, # 완전 random sampling
    max_length=50,
    top_k=0 # w/o top_k 추출
)

print(tokenizer.decode(sample_output.tolist()[0], skip_special_tokens=True))
```

```
이순신은 조선 중기의 무신이다.</s><s> 고왕목(高王玉, 리빙프라자효림한방병원 이사장) 별세: 보혜
```

Sampling-based

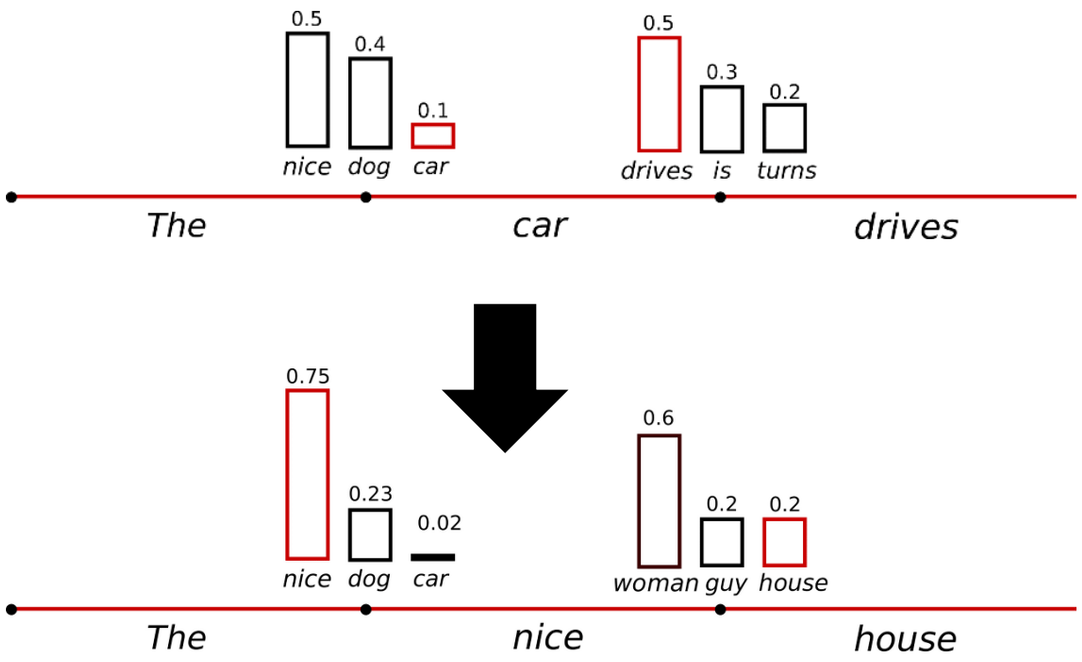
2. Temperature

- Sampling with Temperature

- “ t ∈ [0, 1) “ skews the distribution towards high probability events

  -> implicitly lowers the mass in the tail distribution.

- lowering the temperature improves generation quality, it comes at the cost of decreasing diversity (Caccia et al., 2018; Hashimoto et al., 2019).



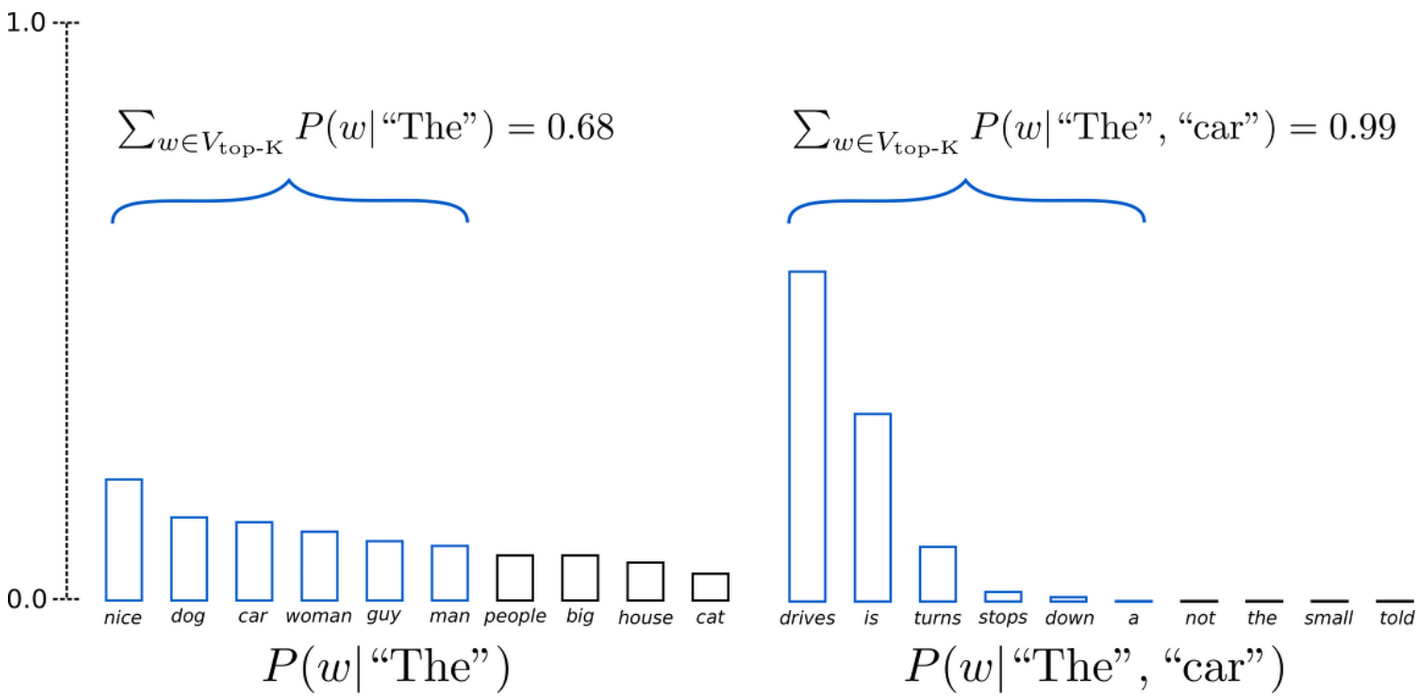

```
# use temperature to decrease the sensitivity to low probability candidates
sample_output = model.generate(
    input_ids,
    do_sample=True,
    max_length=50,
    top_k=0,
    temperature=0.7
    )

print(tokenizer.decode(sample_output.tolist()[0], skip_special_tokens=True))
```

```
이순신은 조선 중기의 무신이다.</s><s> 그는 “다섯세트는 에로틱하고 무장하고 있어
```

Sampling-based

3. Top-K sampling

- Top-K sampling
- a popular alternative sampling procedure (Fan et al., 2018; Holtzman et al., 2018; Radford et al., 2019)
- At each time step, the top k possible next tokens are sampled from according to their relative probabilities




```
# set top_k to 50
sample_output = model.generate(
    input_ids,
    do_sample=True,
    max_length=50,
    top_k=50
)

print(tokenizer.decode(sample_output.tolist()[0],
                       skip_special_tokens=True))
```

이순신은 조선 중기의 무신이다.</s><s> 17일 방송된 KBS 2T

Sampling-based

3. Top-K sampling

- Difficulty in choosing a suitable value of k

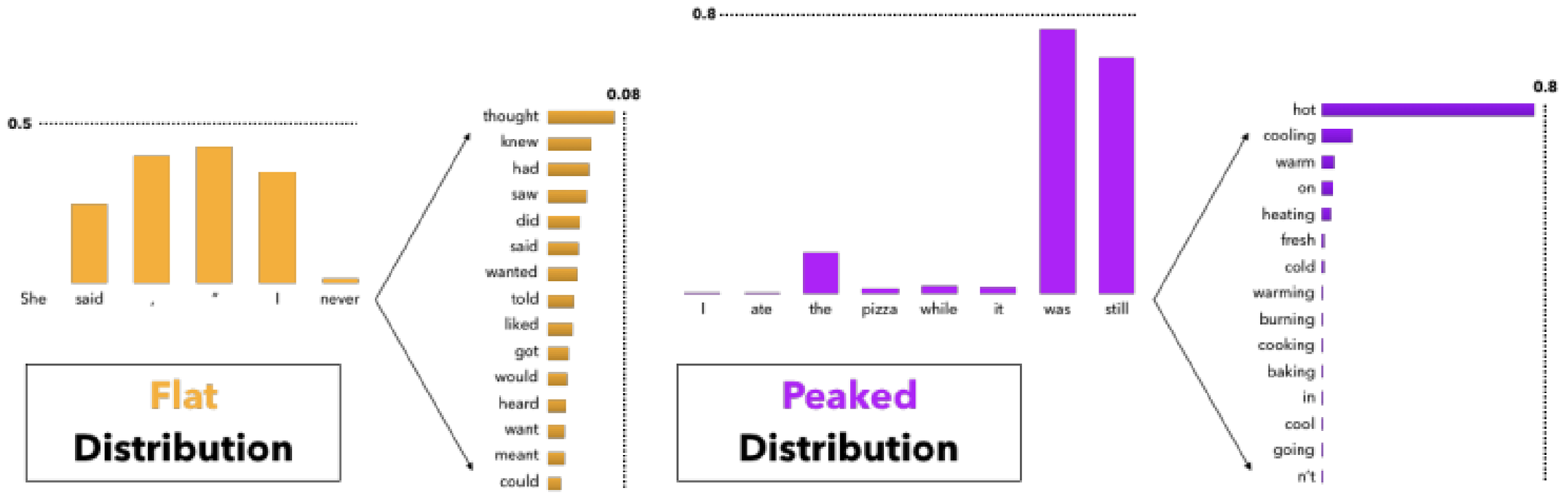
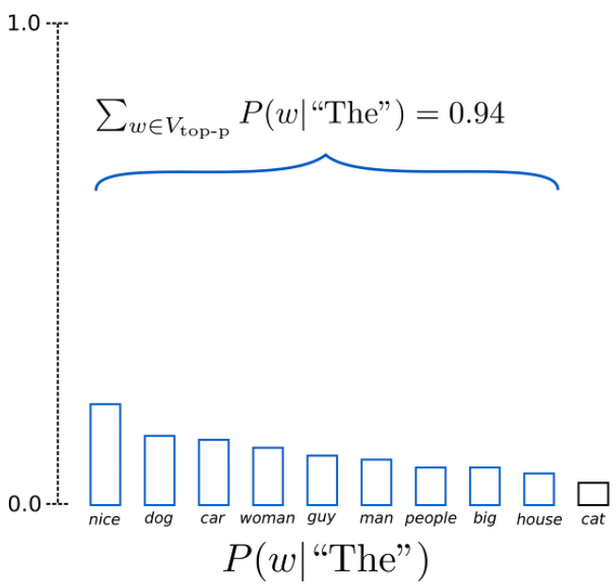


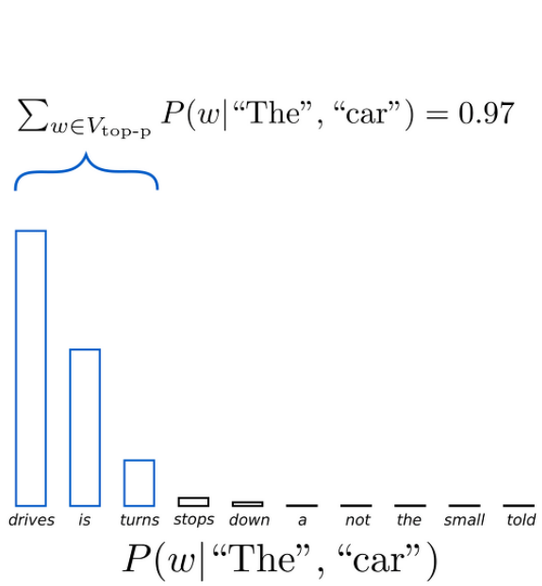
Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Sampling-based

4. Top-p sampling

- Top-p sampling(Nucleus Sampling)
- a new stochastic decoding method
- key idea: use the shape of the probability distribution to determine the set of tokens to be sampled from.






```
# set top_k = 50 and set top_p = 0.95 and num_return_sequences = 3
sample_outputs = model.generate(
    input_ids,
    do_sample=True,
    max_length=50,
    top_k=20,
    top_p=0.92,
    num_return_sequences=3
)

print("Output:₩n" + 100 * '-')
for i, sample_output in enumerate(sample_outputs):
  print("{}: {}".format(i, tokenizer.decode(sample_output.tolist(), skip_special_tokens=True)))
```

```
Output:
----------------------------------------------------------------------------------------------------
0: 이순신은 조선 중기의 무신이다.</s><s> "내 딸도 못 보내지만 내 아내도 못 보내지만 내 아내도 못
1: 이순신은 조선 중기의 무신이다.</s><s> 하지만, 그는 "내 인생은 나에게 마지막 인생과 같다" 며
2: 이순신은 조선 중기의 무신이다.</s><s> 하지만, 그의 아버지는 그의 어머니인 신응수(申應洙)에 의
```

Text Generation Method

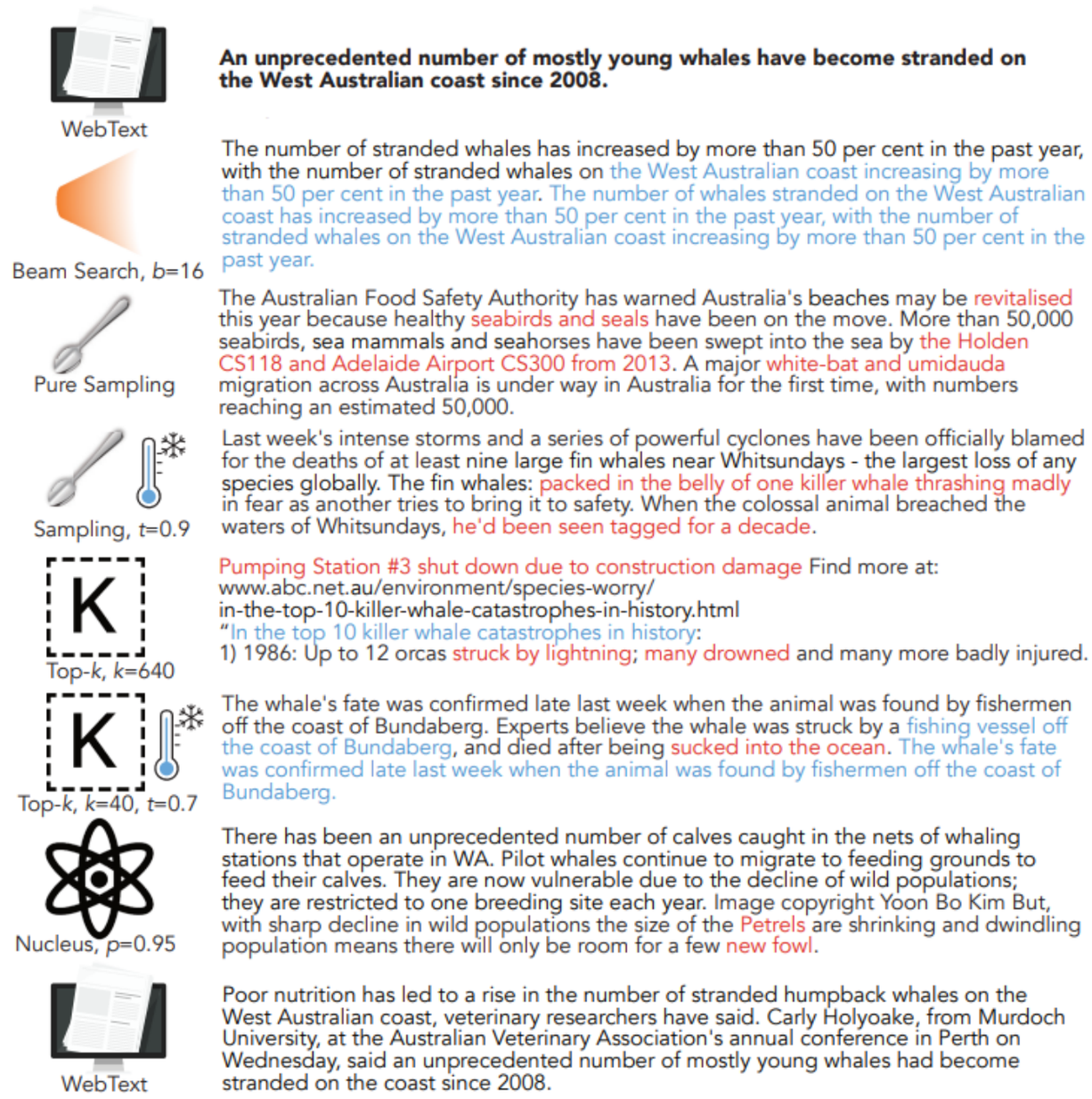


Figure 3: Example generations continuing an initial sentence. Maximization and top-k truncation methods lead to copious repetition (highlighted in blue), while sampling with and without temperature tends to lead to incoherence (highlighted in red). Nucleus Sampling largely avoids both issues.

Evaluation

Perplexity

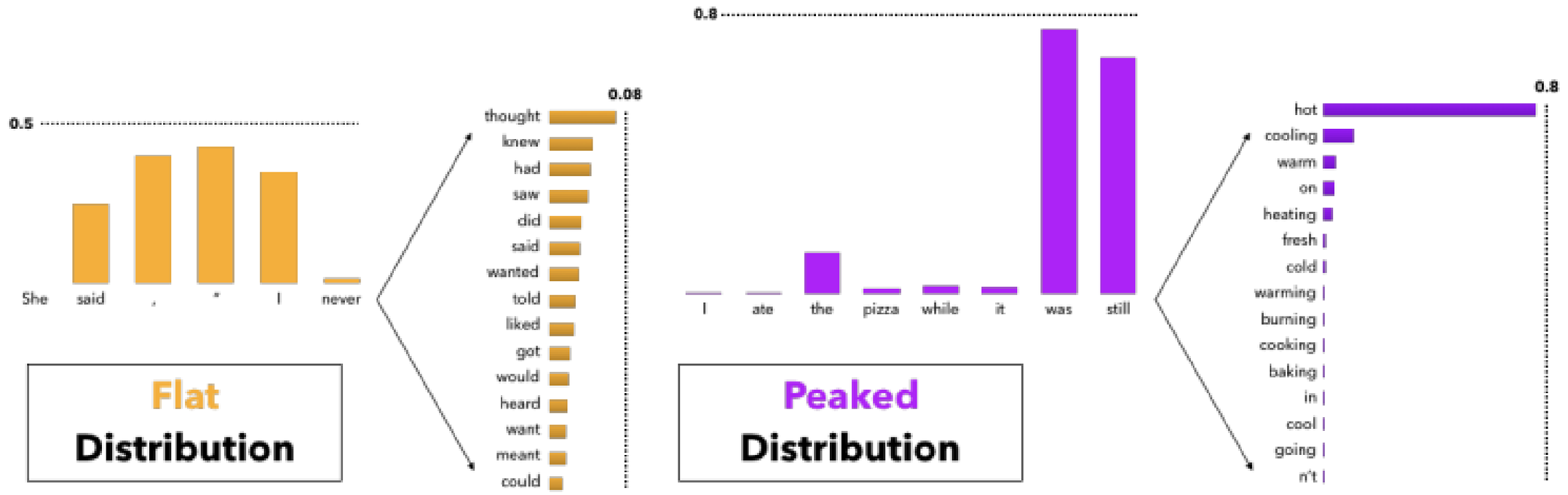


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Evaluation

Perplexity

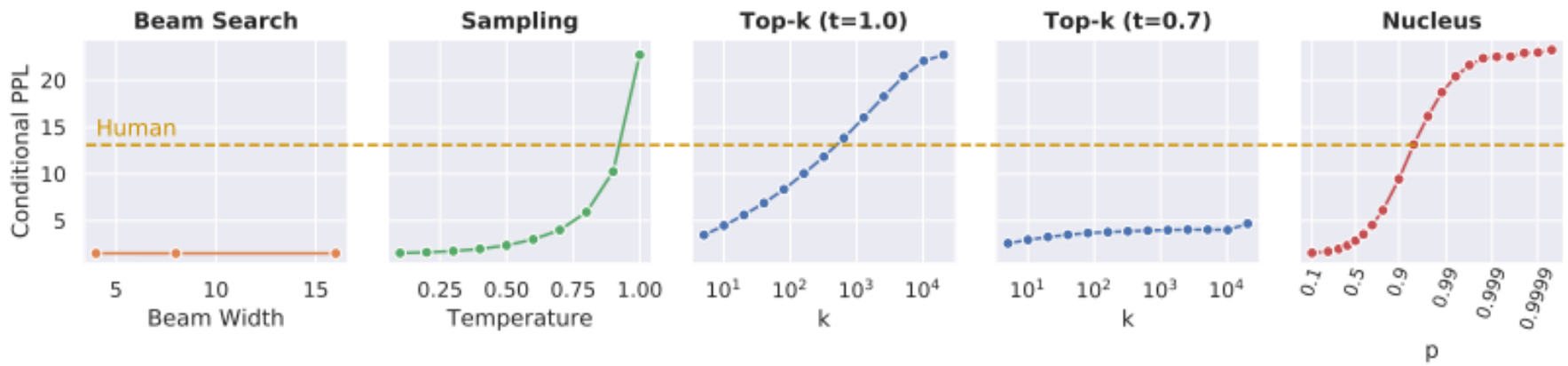


Figure 6: Perplexities of generations from various decoding methods. Note that beam search has unnaturally low perplexities. A similar effect is seen using a temperature of 0.7 with top-k as in both Radford et al. (2019) and Fan et al. (2018). Sampling, Top-k, and Nucleus can all be calibrated to human perplexities, but the first two face coherency issues when their parameters are set this high.

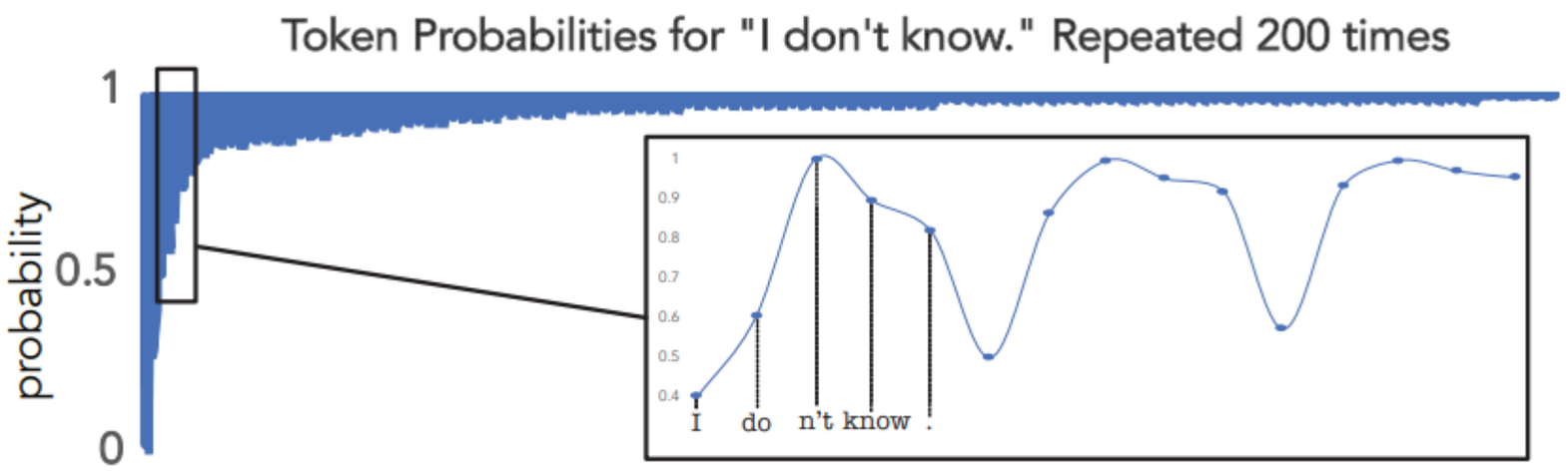


Figure 4: The probability of a repeated phrase increases with each repetition, creating a positive feedback loop. We found this effect to hold for the vast majority of phrases we tested, regardless of phrase length or if the phrases were sampled randomly rather than taken from human text.

Evaluation

Perplexity

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**

Table 1: Main results for comparing all decoding methods with selected parameters of each method. The numbers *closest to human scores* are in **bold** except for HUSE (Hashimoto et al., 2019), a combined human and statistical evaluation, where the highest (best) value is **bolded**. For Top-k and Nucleus Sampling, HUSE is computed with interpolation rather than truncation (see §6.1).

Evaluation

Natural Langugage does not Maximize probability

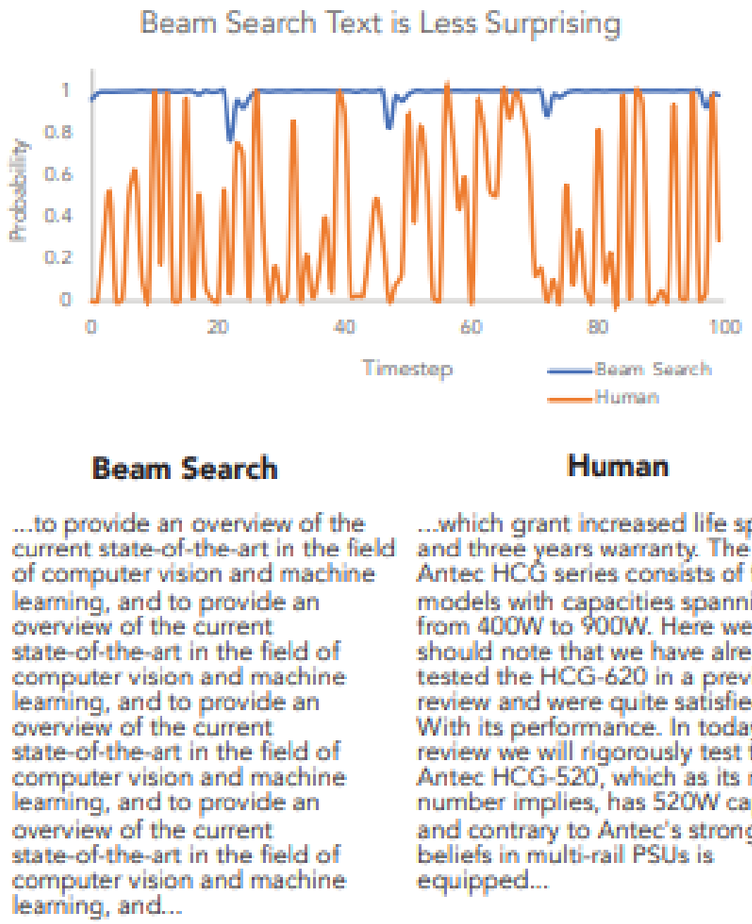


Figure 2: The probability assigned to tokens generated by Beam Search and humans, given the same context. Note the increased variance that characterizes human text, in contrast with the endless repetition of text decoded by Beam Search.

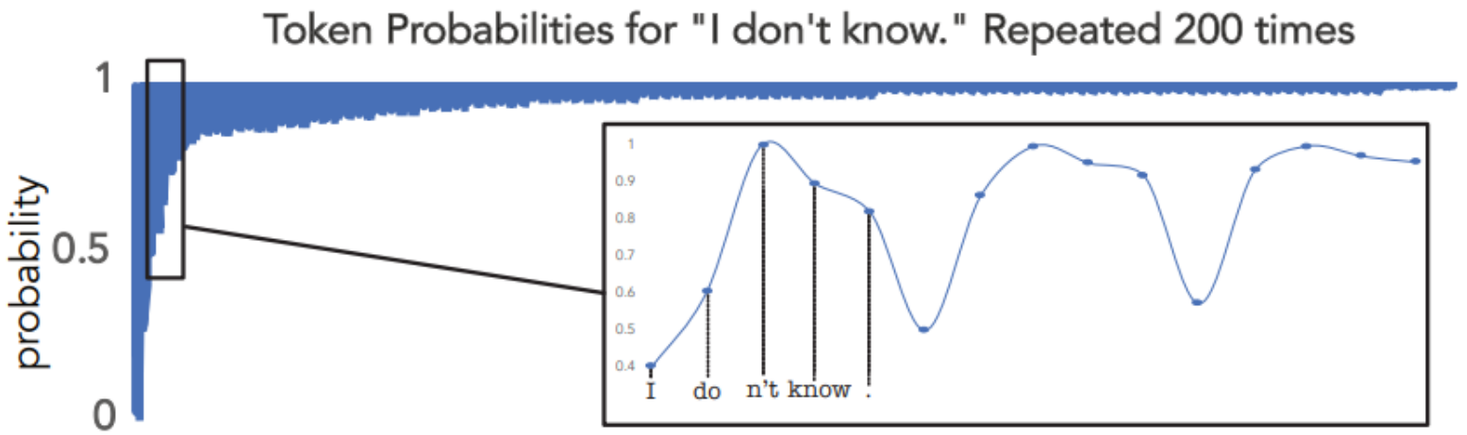


Figure 4: The probability of a repeated phrase increases with each repetition, creating a positive feedback loop. We found this effect to hold for the vast majority of phrases we tested, regardless of phrase length or if the phrases were sampled randomly rather than taken from human text.

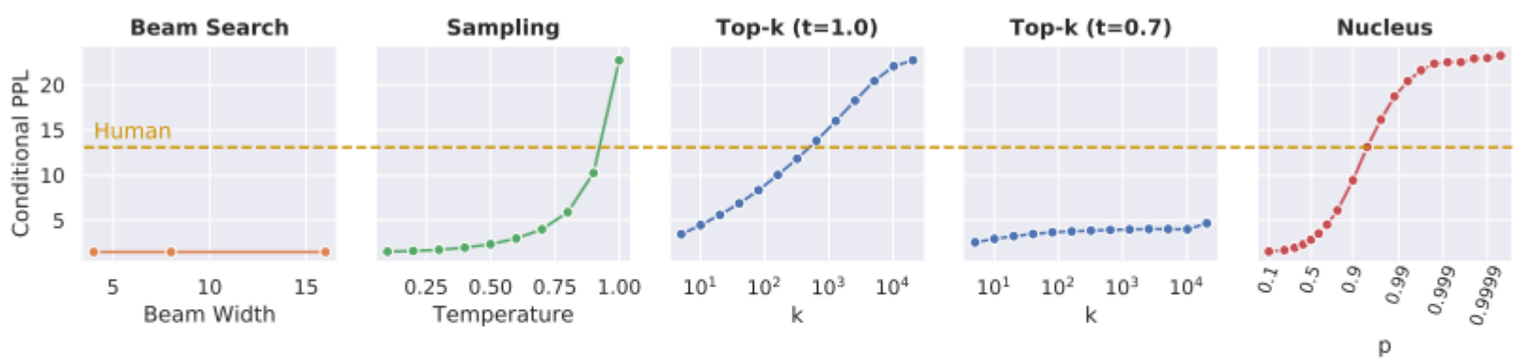


Figure 6: Perplexities of generations from various decoding methods. Note that beam search has unnaturally low perplexities. A similar effect is seen using a temperature of 0.7 with top-k as in both Radford et al. (2019) and Fan et al. (2018). Sampling, Top-k, and Nucleus can all be calibrated to human perplexities, but the first two face coherency issues when their parameters are set this high.

Evaluation

Natural Langugage does not Maximize probability

- Why human-written text is not the most probable text?
- Grice’s Maxims of Communication
  - people optimize against stating the obvious
  - making every word as predictable as possible will be disfavored

Distributional Statistical Evaluation
Zipf Distribution Analysis

- Zipf’s law

- exponential relationship between the rank of a word and its frequency in text.

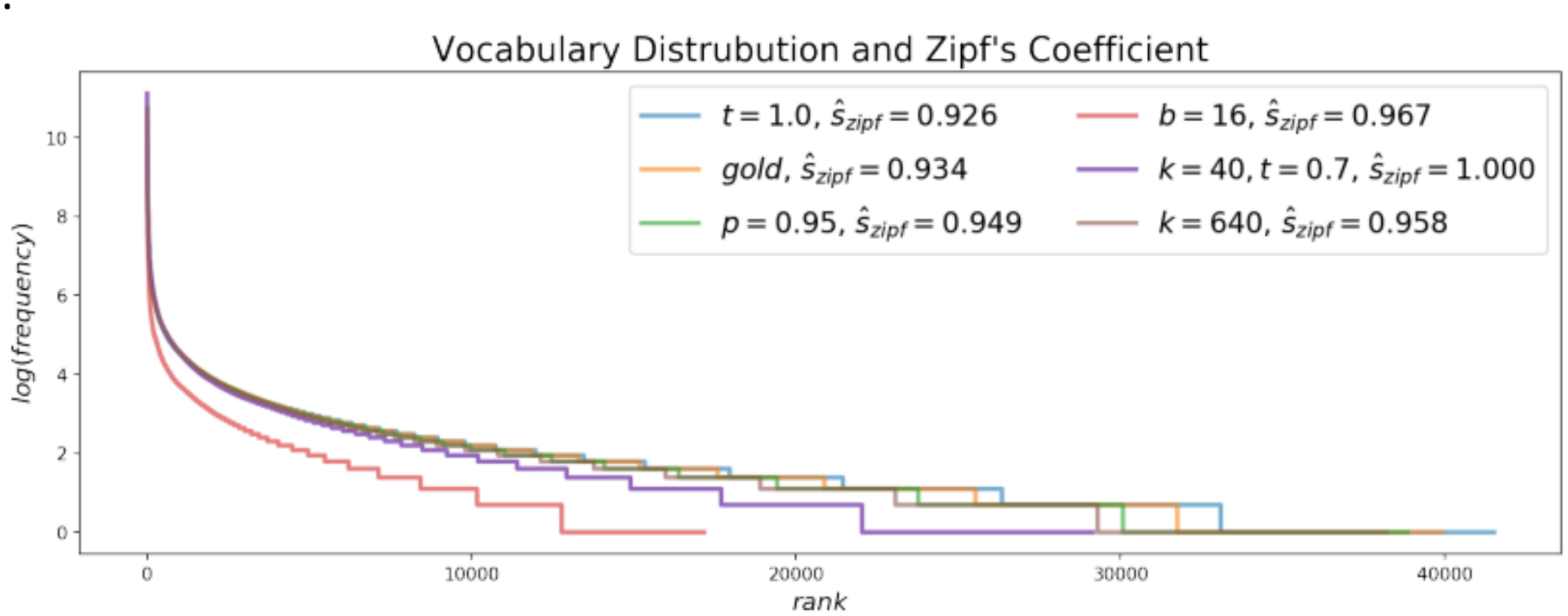


Figure 7: A rank-frequency plot of the distributional differences between n-gram frequencies of human and machine text. Sampling and Nucleus Sampling are by far the closest to the human distribution, while Beam Search clearly follows a very different distribution than natural language.

Distributional Statistical Evaluation
Self-BLEU

- Self-BLEU (Zhu et al., 2018)
- as a metric of diversity
- lower Self-BLEU score implies higher diversity.

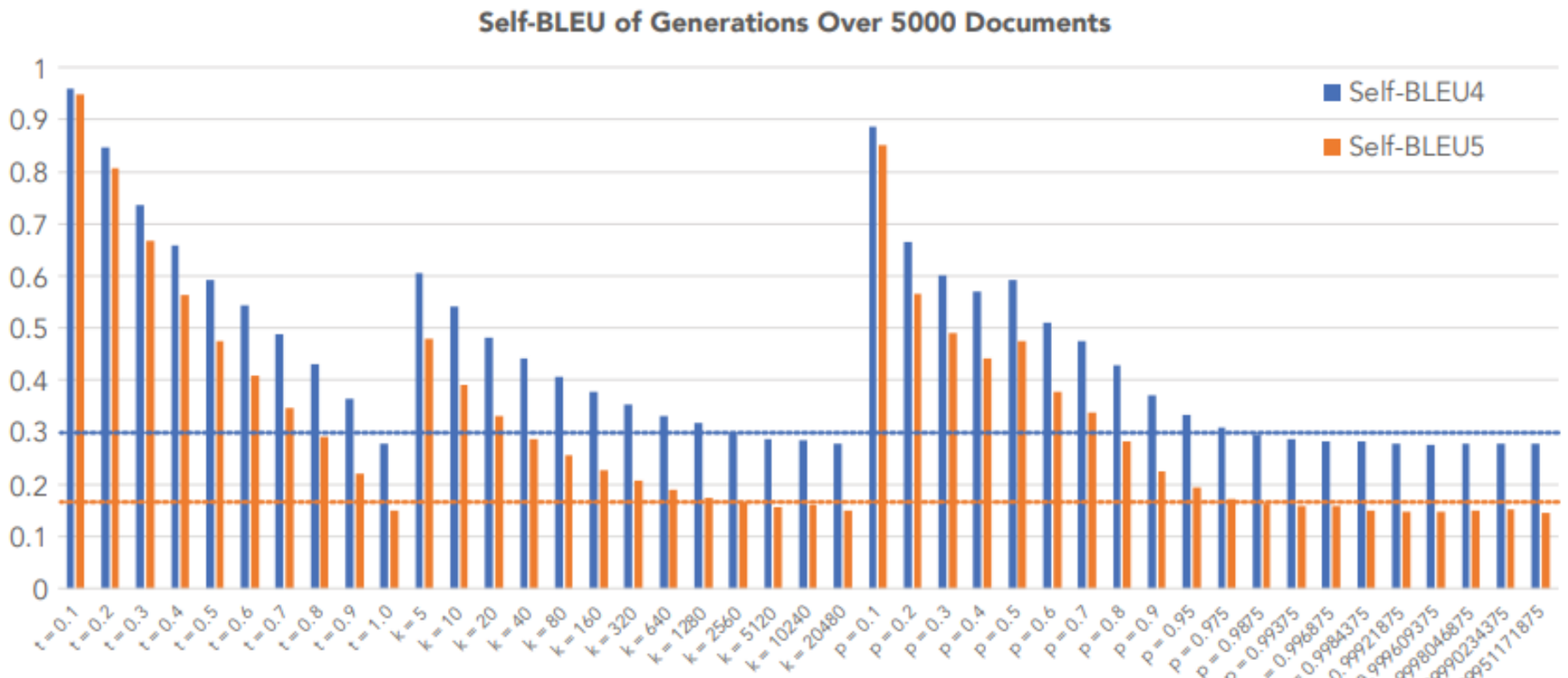


Figure 8: Self-BLEU calculated on the unconditional generations produced by stochastic decoding methods; lower Self-BLEU scores imply higher diversity. Horizontal blue and orange lines represent human self-BLEU scores. Note how common values of $t \in [0.5, 1]$ and $k \in [1, 100]$ result in high self-similarity, whereas "normal" values of $p \in [0.9, 1)$ closely match the human distribution of text.

Distributional Statistical Evaluation

Repetition

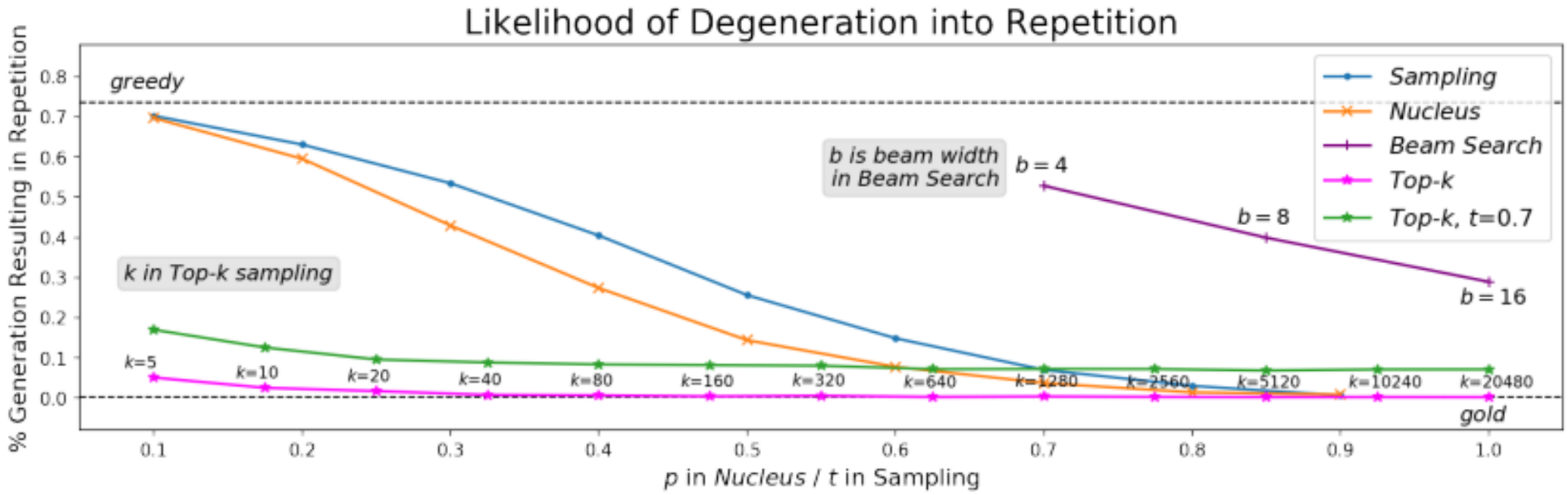


Figure 9: We visualize how often different decoding methods get "stuck" in loops within the first 200 tokens. A phrase (minimum length 2) is considered a repetition when it repeats at least **three** times at the *end* of the generation. We label points with their parameter values except for t and p which follow the x-axis. Values of k greater than 100 are rarely used in practice and values of p are usually in $[0.9, 1)$; therefore Nucleus Sampling is far closer to the human distribution in its usual parameter range. Sampling with temperatures lower than 0.9 severely increase repetition. Finally, although beam search becomes less repetitive according to this metric as beam width increases, this is largely because average length gets shorter as b increases (see Appendix A).

Human Evaluation

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**

Table 1: Main results for comparing all decoding methods with selected parameters of each method. The numbers *closest to human scores* are in **bold** except for HUSE (Hashimoto et al., 2019), a combined human and statistical evaluation, where the highest (best) value is **bolded**. For Top-k and Nucleus Sampling, HUSE is computed with interpolation rather than truncation (see §6.1).

Summary and Take-home message

- Maximization based Method (repetition loop, incoherence problem)
  -> Sampling Method

- Top-K sampling -> Nucleus Sampling(Top-p sampling)


DSAIL
Data Science Artificial Intelligence Lab.