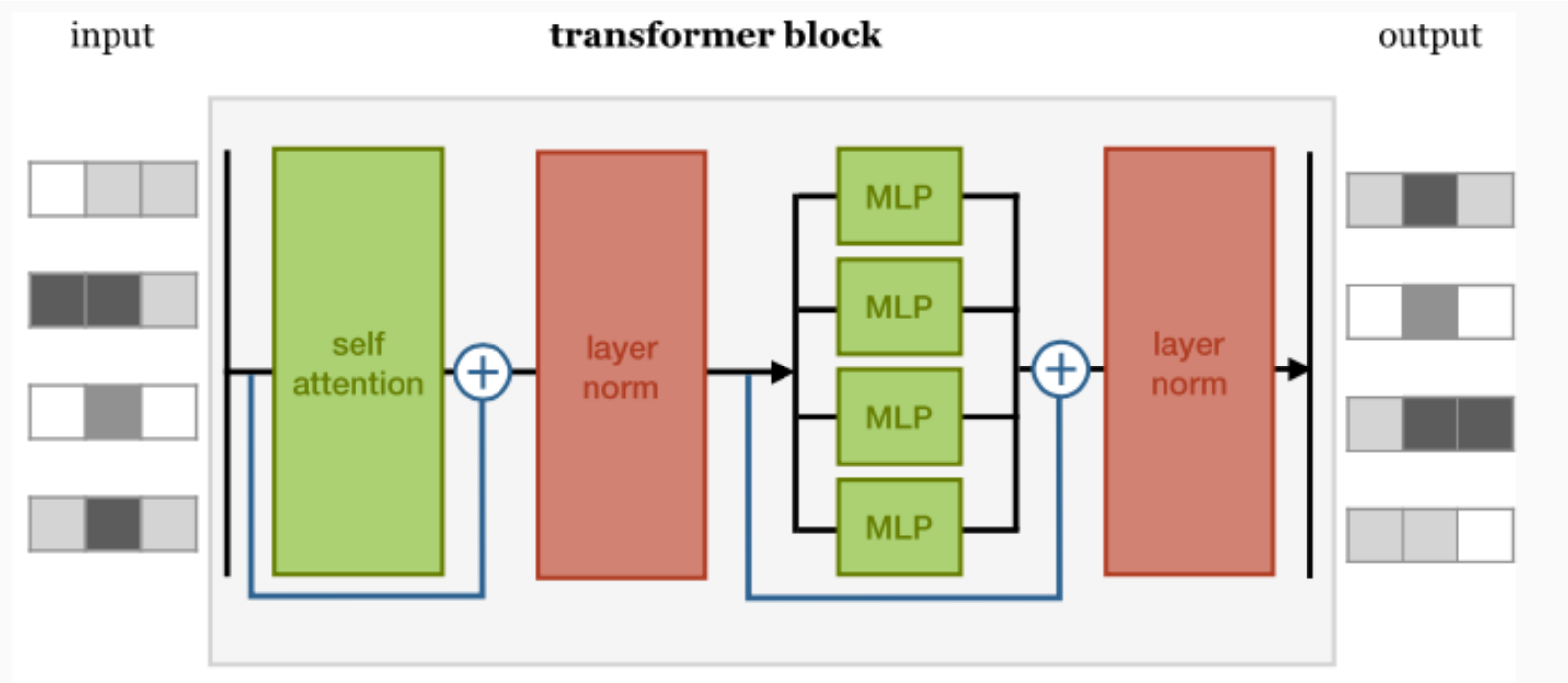
NLU+: Lecture 10
Decoding with LLMs

Shay Cohen
partially based on material from Mohit Iyyer

February 3, 2025

School of informatics

The Story So Far



Language modelling

A language model assigns probability to a sequence $w_1, \ldots, w_n$

We want to use LMs, for example, through prompting

To do that, we need to be able to generate a **continuation**

A note about decoding

Origin of reference to "decoding" is in information theory

> *One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.*

– Warren Weaver, Letter to Norbert Wiener, March 4, 1947

Decoding

Question: How do we generate the most probable continuation?

(Do we want to do that?)

We want to find

$$\arg \max_{w_1,\ldots,w_n} p(w_1, \ldots, w_n \mid \text{prefix})$$

Decoding

Question: How do we generate the most probable continuation?

(Do we want to do that?)

We want to find

$$\arg \max_{w_1,\ldots,w_n} p(w_1, \ldots, w_n \mid \text{prefix})$$

Food for Thought: Could it be done by enumerating all possible continuations?

Decoding

Question: How do we generate the most probable continuation?

(Do we want to do that?)

We want to find

$$\arg \max_{w_1,\ldots,w_n} p(w_1, \ldots, w_n \mid \text{prefix})$$

Food for Thought: Could it be done by enumerating all possible continuations?
Food for Thought: Could it be done by generating word by word?

Decoding

Question: How do we generate the most probable continuation?

(Do we want to do that?)

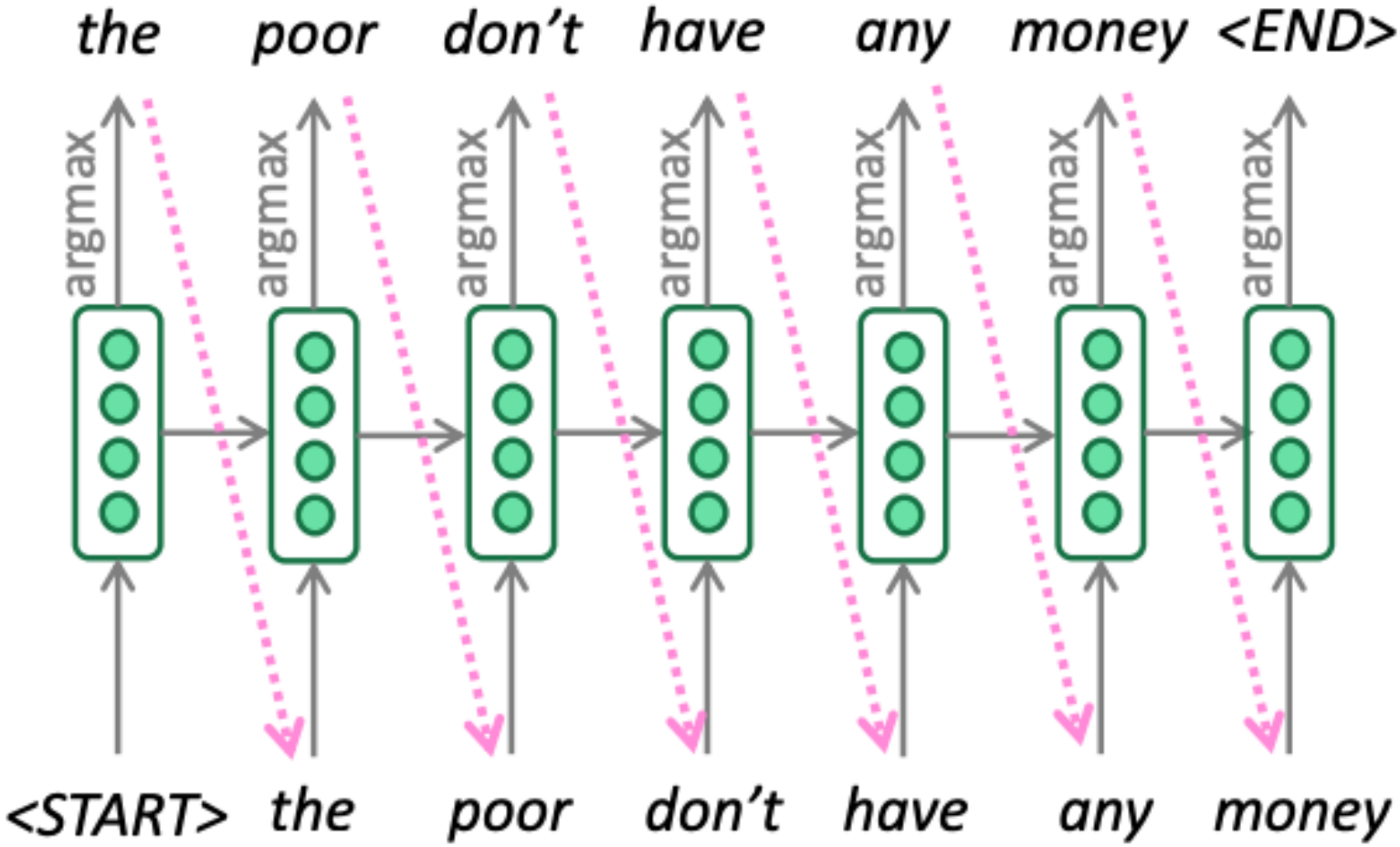
We want to find

$$\arg \max_{w_1,\ldots,w_n} p(w_1, \ldots, w_n \mid \text{prefix})$$

Food for Thought: Could it be done by enumerating all possible continuations?
Food for Thought: Could it be done by generating word by word?

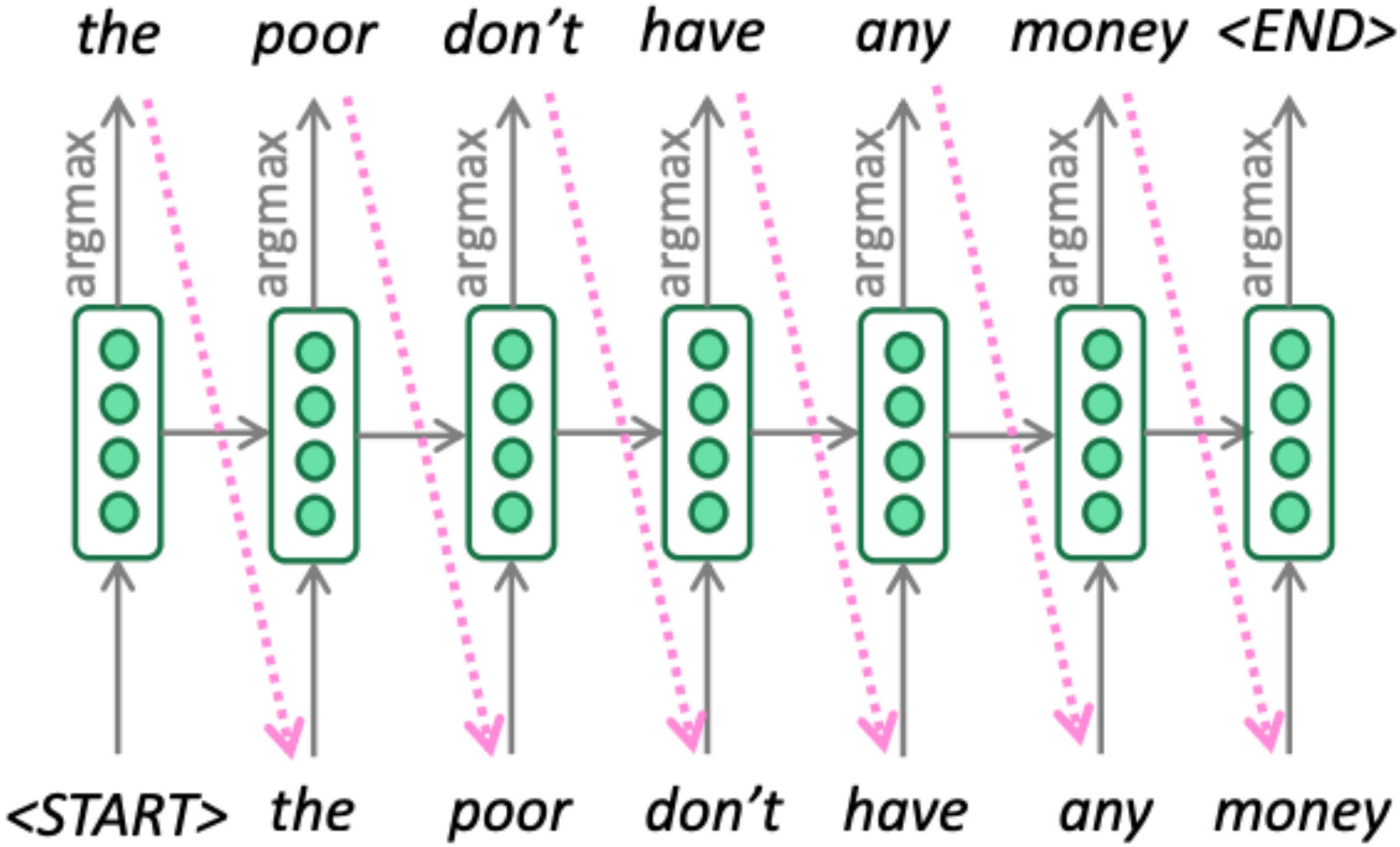
$$p(w \mid context) = \exp\left(f(w, context)\right)/Z$$

Greedy decoding



(figure from Mohit Iyyer)

Greedy decoding



(figure from Mohit Iyyer)

Food for thought: Why not greedy decoding?

Issues with greedy decoding

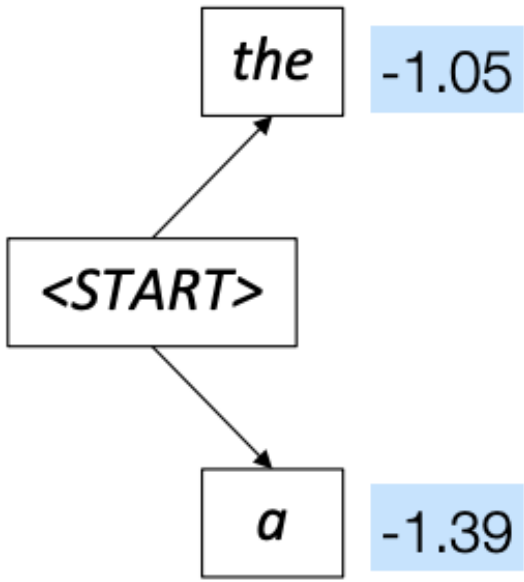
We keep only one continuation at a time

If we make a "mistake", we cannot backtrack and change it

Use instead beam search (just like with parsing from ANLP):

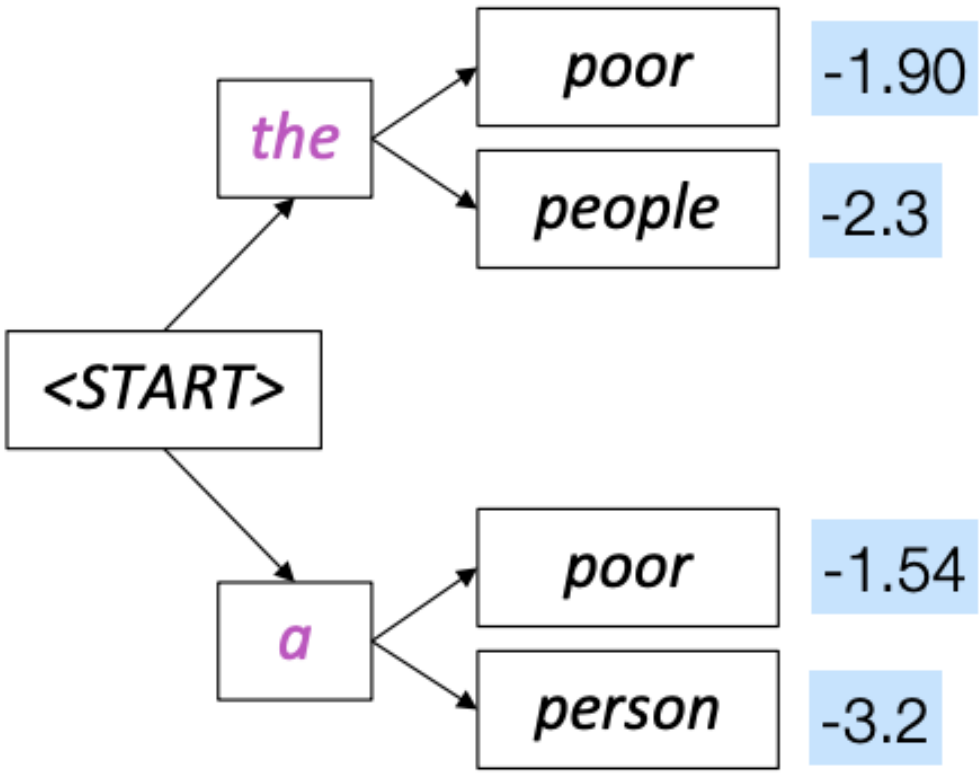
- Retain several partial continuations
- Expand the most promising one at a time

Beam decoding

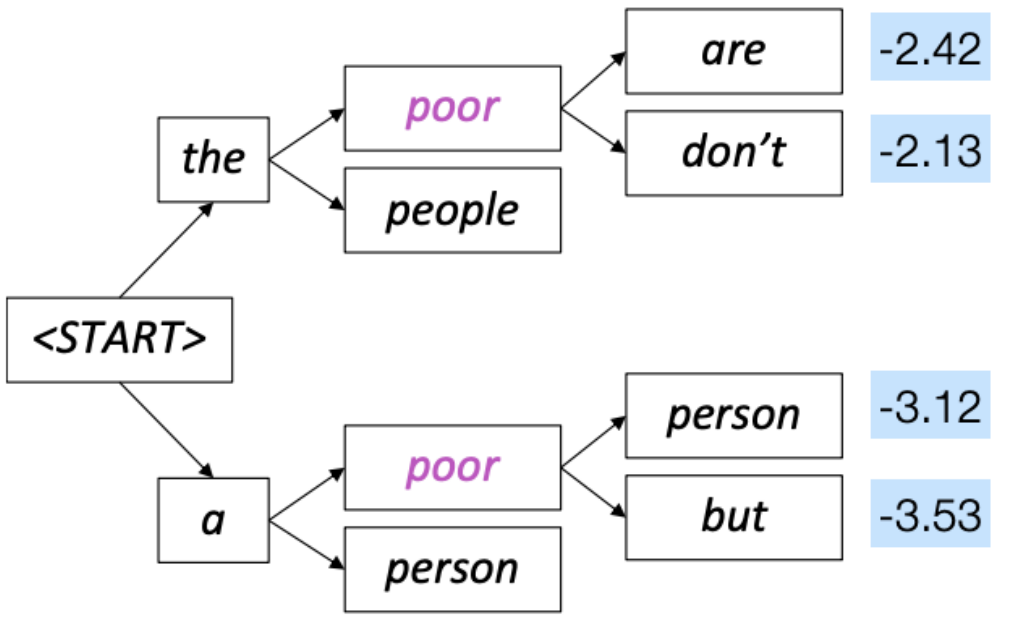


(figures from Mohit Iyyer)

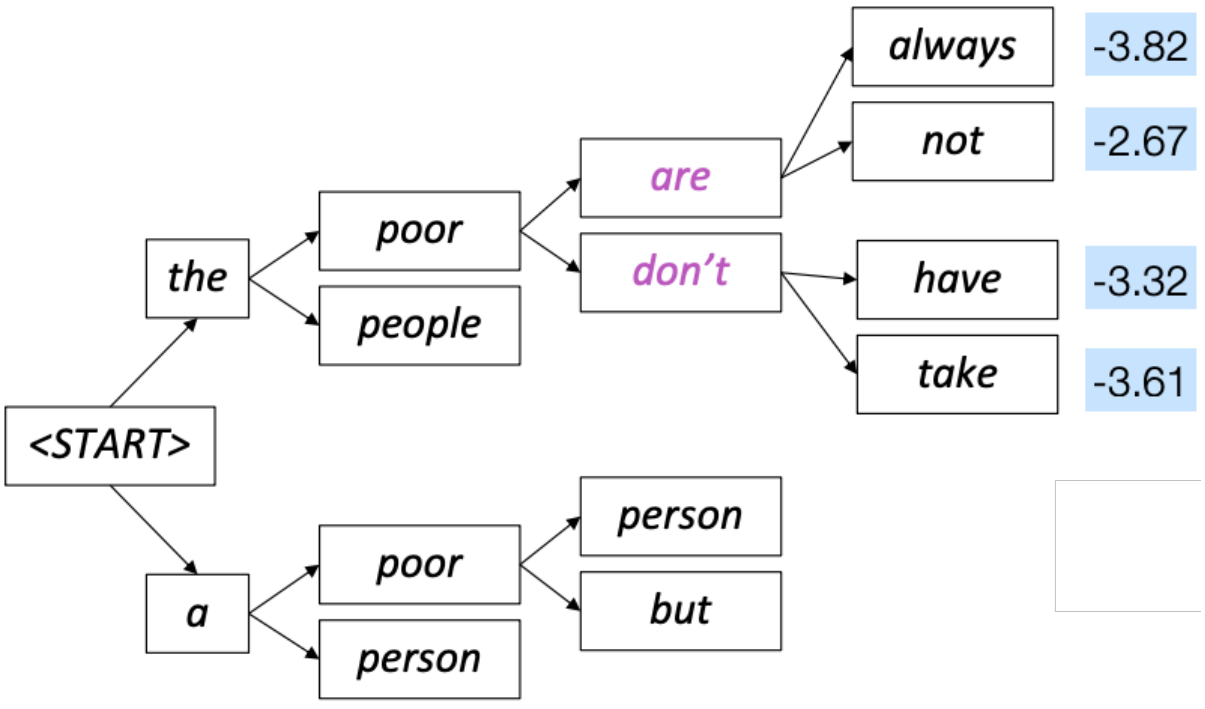
Beam decoding



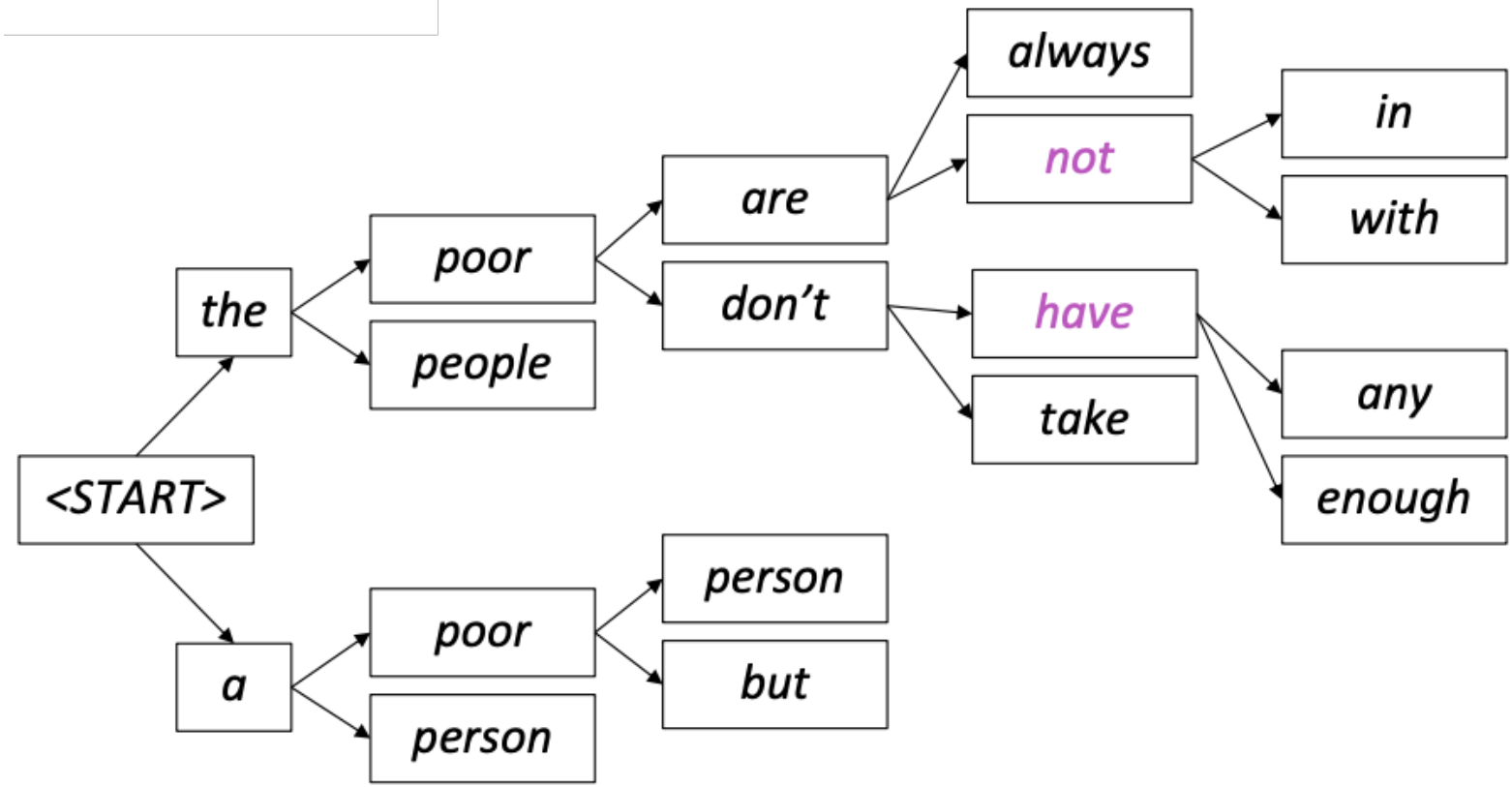
Beam decoding



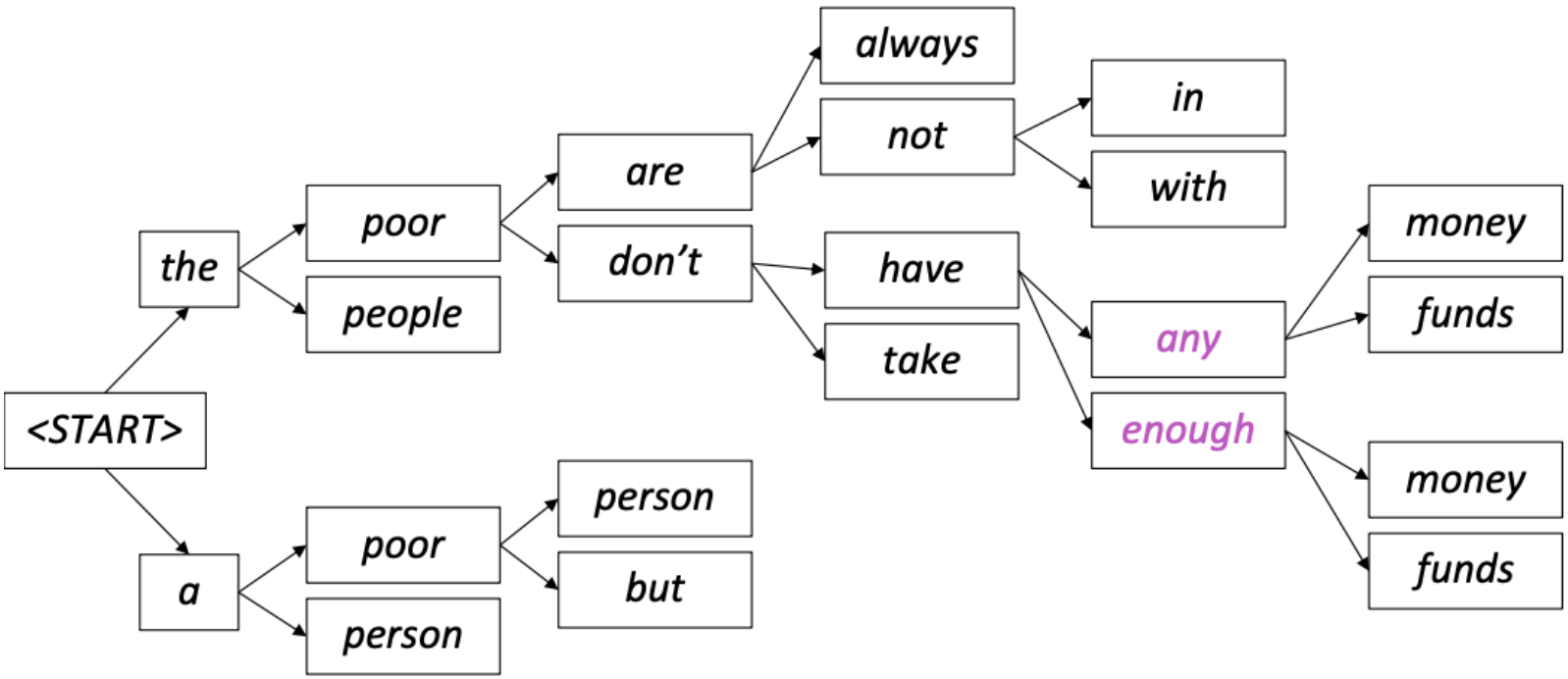
Beam decoding



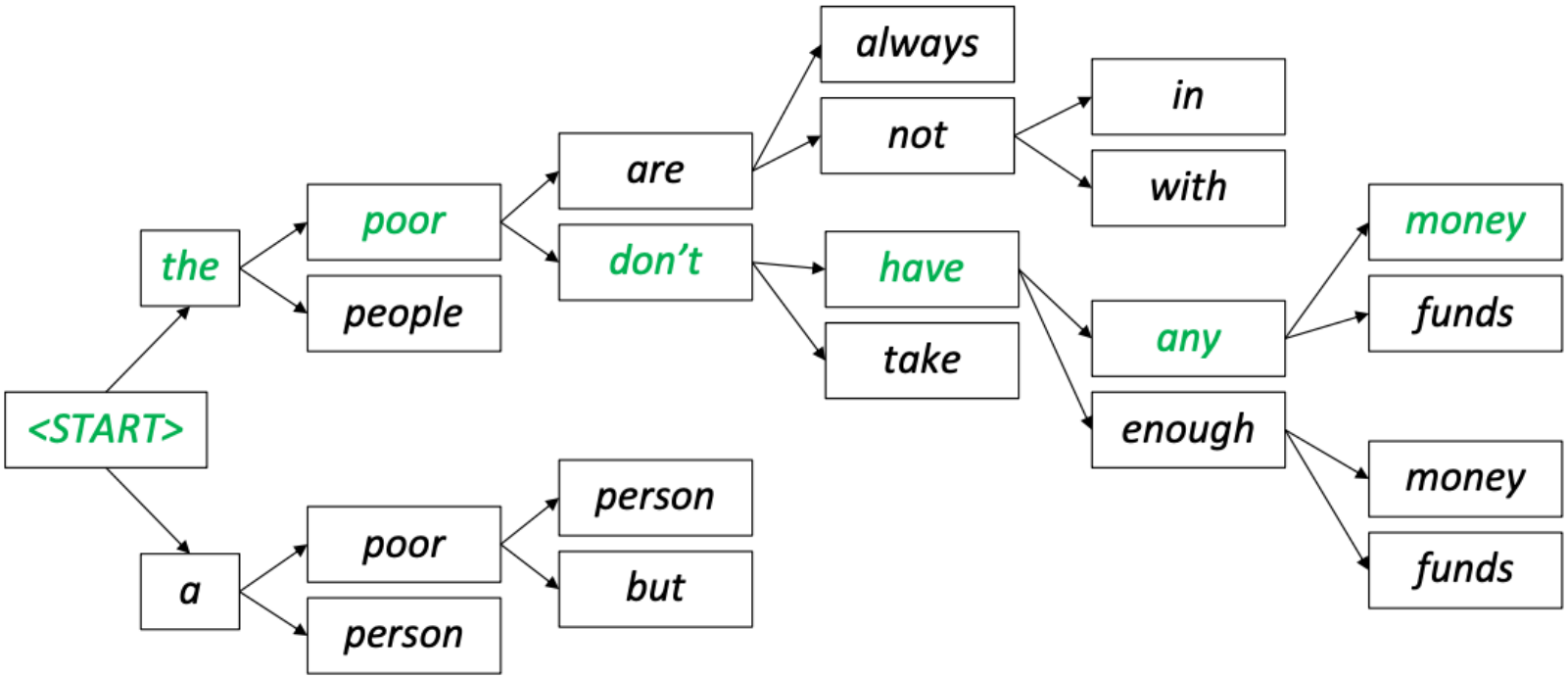
Beam decoding



Beam decoding



Beam decoding



Beam decoding (taken from Kasai et al., 2022)

FCFS Beam Decoding with Controlled Patience

k: beam size, M: maximum length, $\mathcal{V}$: Vocabulary
score$(\cdot)$: scoring function, p: patience factor.

$B_0 \leftarrow \{\langle 0, \text{BOS}\rangle\}$, $F_0 \leftarrow \varnothing$
for $t \in \{1, \ldots, M-1\}$:
  $H \leftarrow \varnothing$, $F_t \leftarrow F_{t-1}$
  for $\langle s, \mathbf{y}\rangle \in B_{t-1}$: # Expansion.
    for $y \in \mathcal{V}$:
      $s \leftarrow \text{score}(\mathbf{y} \circ y)$, H.add$(\langle s, \mathbf{y} \circ y\rangle)$
  $B_t \leftarrow \varnothing$
  while $|B_t| < k$: # Find top k w/o EOS from H.
    $\langle s, \mathbf{y}\rangle \leftarrow H$.max()
    if $\mathbf{y}$.last() $=$ EOS :
      F_t.add$(\langle s, \mathbf{y}\rangle)$ # Finished hypotheses.
    else B_t.add$(\langle s, \mathbf{y}\rangle)$
    if $|F_t| \geq k \cdot p$: # Originally, $p=1$.
      return F_t.max()
    H.remove$(\langle s, \mathbf{y}\rangle)$
return F_t.max()

- F_t - finished hypotheses; B_t - beam of continuing sequences; H - expanded hypotheses before the top-k operation

Implementation of a beam

Need a data structure such as a priority queue

Keep in the priority queue at most k analyses

When coming to decode the next word, pop from the queue, expand, and push back

Priority queues can typically be made efficient so that the computational complexity of their operations is no more than $O(\log k)$ or even $O(1)$

The effect of beam size k

If k is small, we have similar problems to greedy decoding ($k = 1$ *is* greedy decoding), and we will make many mistakes

A larger k has the following issues:

- May be computationally expensive
- It has been shown that increasing k too much can lead to worse translations, for example (Tu et al., 2017; Koehn et al., 2017)
- There is a strong bias with larger k to generate short translations
- (Note that in general a continuation which is short is apriori preferred. Why?)

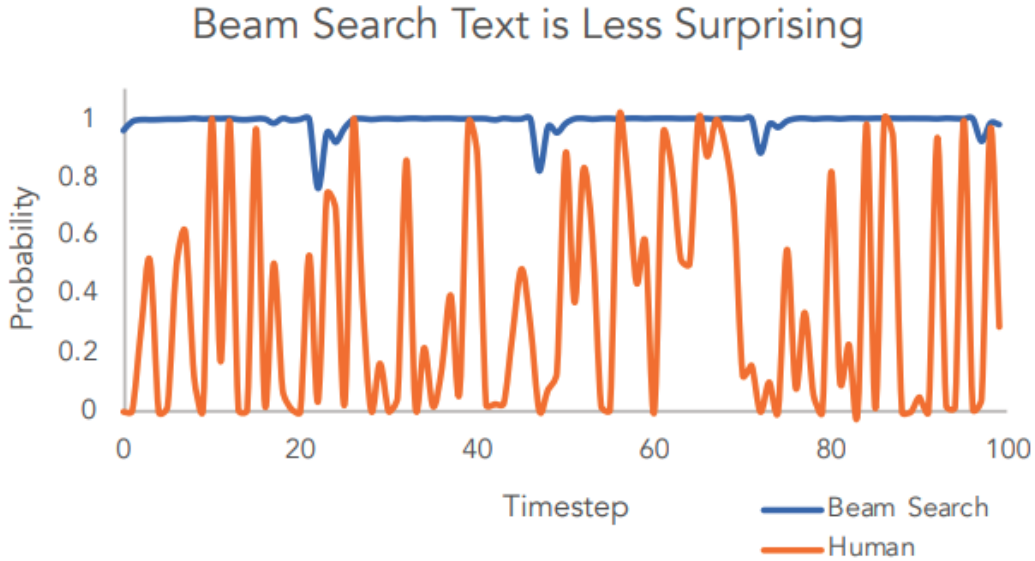
Overall, needs to find the "right" k as a hyperparameter

Sampling to the rescue

Why sampling?

- Diversity in output - generate varied responses
- Avoiding repetition
- Exploring alternatives to the default maximal probability
- Sounding more natural

The Story So Far



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...

Sampling decoding methods

- At each step, sample from the probability distribution over words rather than choosing the highest probability word
- No multiple hypotheses

Other sampling methods:

- Sample only from the top k most probable words
  Food for thought: Why truncating the less probable words?

Sampling decoding methods

- At each step, sample from the probability distribution over words rather than choosing the highest probability word
- No multiple hypotheses

Other sampling methods:

- Sample only from the top k most probable words
  Food for thought: Why truncating the less probable words? Heavy tail
- $k = 1$ is greedy search, for $k = \infty$, we get sampling as above
- Larger k: more diverse output; Smaller k: generic/safe output
- (This parameter can be controlled in most APIs for LLMs!)

Nucleus sampling (Holtzman et al., 2020)

Rather than choosing the top k tokens as an option, truncate the tail of the next-token probability distribution

Choice of available tokens is now dynamic and depends on the distribution at a specific time step

Balances between risky output (in the tail of the distribution) and generic output (in the head of the distribution)

Temperature in LLMs

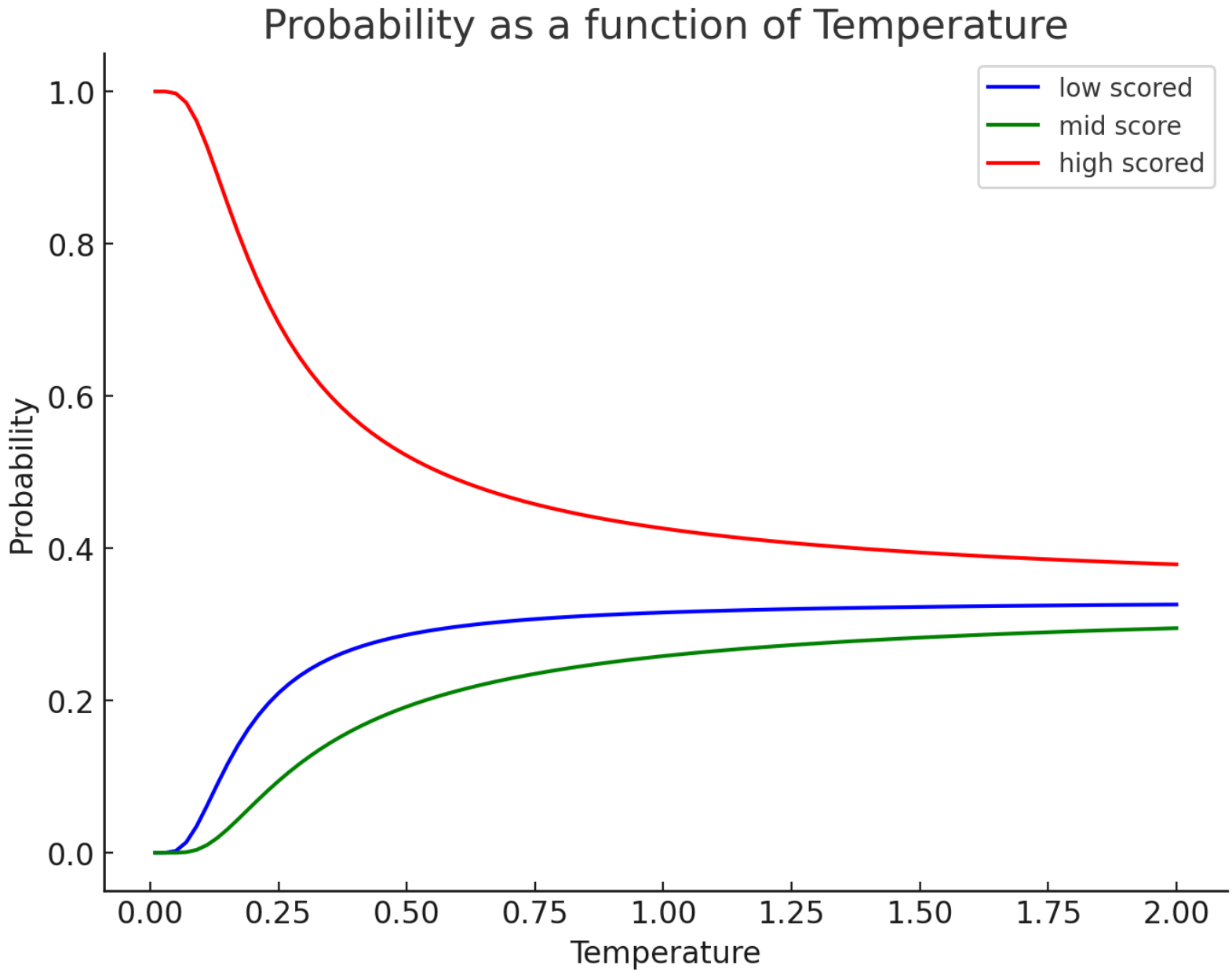
The higher the temperature τ is, the more "random" the output is.

Temperature, effectively, directly changes the softmax distribution:

$$\frac{\exp\left(score(w)/\tau\right)}{\sum_{w'} \exp\left(score(w')/\tau\right)}$$

Lower temperature - make more determinstic choices based on the most probable tokens

Temperature plot



Holtzman et al. (2020)

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

Holtzman et al. (2020)

WebText

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

Beam Search, b=16

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.

Holtzman et al. (2020)

WebText

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

Beam Search, b=16

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.

Pure Sampling

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.

Holtzman et al. (2020)

WebText

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

Beam Search, b=16

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.

Pure Sampling

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.

Sampling, t=0.9

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.

Holtzman et al. (2020)

WebText

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

Beam Search, b=16

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.

Pure Sampling

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.

Sampling, t=0.9

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.

Top-k, k=640

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.

Holtzman et al. (2020)

WebText

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

Beam Search, b=16

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.

Pure Sampling

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.

Sampling, t=0.9

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.

Top-k, k=640

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.

Top-k, k=40, t=0.7

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.

Holtzman et al. (2020)

Sampling, t=0.9

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.

Top-k, k=640

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.

Top-k, k=40, t=0.7

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.

Nucleus, p=0.95

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.

From the paper: "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 issue"

Holtzman et al. (2020) - another example

WebText

So what's new in my life?
09/11/18 - Just got back from vacation.

Beam Search, b=16

Just got back from vacation.
09/10/18 - Just got back from vacation. Just got back from vacation.
09/09/18 - Just got back from vacation. Just got back from vacation.
09/08/18 - Just got back from vacation. Just got back from vacation.

Pure Sampling

01/05/19 - Visited The Subaru Canada Heritage15th Star for my vehicle research. That was a wonderful experience. Thanks again for your knowledge and support.
05/23/19 - I was thrilled to receive my Subaru Legacy character! It fits me perfectly and makes me feel like a kid again!
06/15/19 - Just regained my licence for my different coloured 2016 WRX / STI...

Sampling, t=0.9

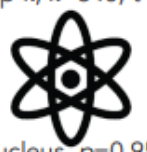
Just got back from vacation.
09/07/18 - Happy 9th birthday to me! Happy 9th birthday to me!
09/01/18 - Sofwidden also! Sofwidden also!
08/25/18 - Happy birthday to me! Happy birthday to me!
08/17/18 - Night shift! □□Night shift! □□
08/17/18 - Lucky me! Lucky me!

Top-k, k=640

Just got back from vacation.
09/03/18 - After more than 40 years, I might have an apartment in Virginia Beach. After more than 40 years, I might have an apartment in Virginia Beach.
08/20/18 - Going for a hike at Mount Eerie in Northeast Virginia Spring Going for a hike at Mount Eerie in Northeast Virginia Spring

Top-k, k=640, t=0.7

Just got back from vacation.
09/08/18 - I've gotten really sick. - I've gotten really sick.
09/07/18 - My wife and I are getting married in February.
- My wife and I are getting married in February.
09/06/18 - I'm so excited to go back to college this fall.
- I'm so excited to go back to college this fall.



Nucleus, p=0.95

Just got back from vacation.
07/12/18 - Happy birthday to Swingu, who is nearly 5 years old. I would like to say hi to him on the road as well as when I ride with him. You cannot go to work without feeling physically sick or psychologically exhausted because you can barely breathe. Even if you ride on rollercoaster even once, it is easy to recover from the physical side of it.

WebText

I just got back from a much needed and really great nine day vacation to my remote Arizona property. It was a really restful and relaxing visit. I got a lot accomplished while I was there, but still found time to just goof off and have fun too. I got to do some astronomy, even though the weather was pretty cloudy most of the time. Here is a 50 minute exposure of M101. It turned out pretty good.

Back to temperature

Let's check it out on OpenAI's playground!

`https://platform.openai.com/playground?mode=chat`

Locally typical sampling (Meister et al., 2022)

Main premise behind this decoding algorithm: the information content of human utterances should be "typical"

The information content, therefore, should be close to the *expected* information content

How do we formalise this?

Typicality (Meister et al., 2022)

Consider the probability over next words $p(w_t \mid w_1, \ldots, w_{t-1})$ that a model can generate where the context is $w_1, \ldots, w_{t-1}$.

We can calculate the entropy of this distribution as

$$-\sum_w p(w_t = w \mid \mathit{context}) \log p(w_t = w \mid \mathit{context}).$$

What does $-\log p(w)$ mean?

Food for Thought: Say you wanted have a probability over a collection of words. You wanted to encode the words using bits in some "optimal" way. What would it be?

Typicality (Meister et al., 2022)

Consider the probability over next words $p(w_t \mid w_1, \ldots, w_{t-1})$ that a model can generate where the context is $w_1, \ldots, w_{t-1}$.

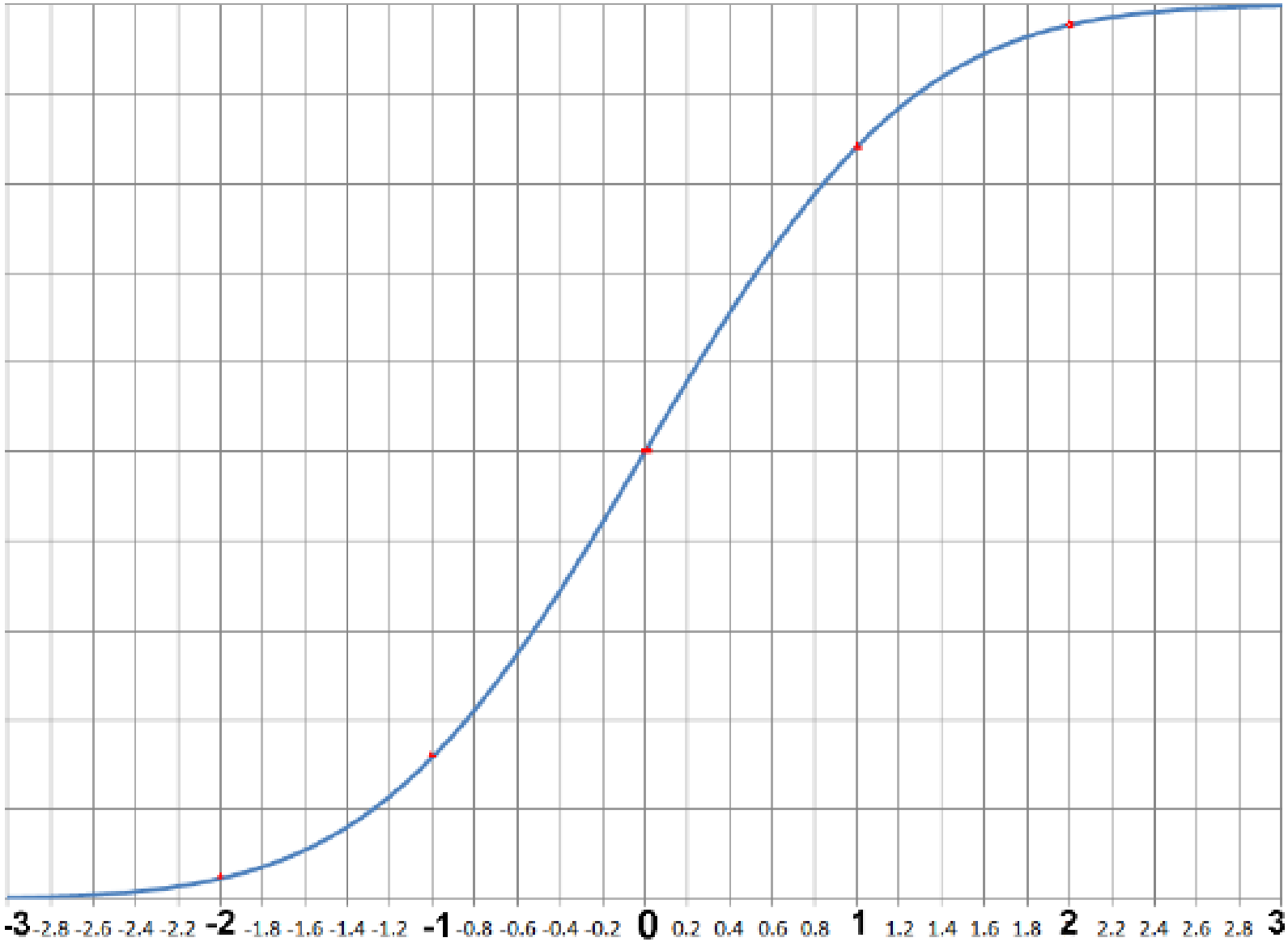
We can calculate the entropy of this distribution as

$$-\sum_w p(w_t = w \mid \mathit{context}) \log p(w_t = w \mid \mathit{context}).$$

What does $-\log p(w)$ mean?

Food for Thought: Say you wanted have a probability over a collection of words. You wanted to encode the words using bits in some "optimal" way. What would it be?

Food for Thought: How does binary search work?


-3 -2.8 -2.6 -2.4 -2.2 -2 -1.8 -1.6 -1.4 -1.2 -1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 2.2 2.4 2.6 2.8 3

Locally typical sampling

The negated log probability is this optimal number of bits you would need to encode the word

The entropy is the average number of symbols you would need to encode a word

Locally typical sampling (Meister et al., 2022): sample only words that are close to the typical number of bits you would need

Locally typical sampling

Define a set C of words that are possible as

$$\arg\min_{C} \sum_{w \in C} \left| H(w_t = w \mid context) + \log p(w_t = w \mid context) \right|$$

$$subject\ to \quad \sum_{w \in C} p(w_t = w \mid context) \geq \tau$$

Locally typical sampling

Define a set C of words that are possible as

$$\arg\min_{C} \sum_{w \in C} \left| H(w_t = w \mid context) + \log p(w_t = w \mid context) \right|$$

$$subject\ to \quad \sum_{w \in C} p(w_t = w \mid context) \geq \tau$$

Food for thought: What if we changed the minimisation problem to $\arg\min_C |C|$ with the same "subject to" constraint?

Abstractive Summarization (CNN/DailyMail)

Prompt	(CNN) The attorney for a suburban New York cardiologist charged in what authorities say was a failed scheme to have another physician hurt or killed is calling the allegations against his client "completely unsubstantiated." Appearing Saturday morning on CNN's "New Day," Randy Zelin defended his client, Dr. Anthony Moschetto, who faces criminal solicitation, conspiracy, burglary, arson, criminal prescription sale and weapons charges in connection to what prosecutors called a plot to take out a rival doctor on Long Island. "None of anything in this case has any evidentiary value," Zelin told CNN's Christi Paul...
Reference	A lawyer for Dr. Anthony Moschetto says the charges against him are baseless. Moschetto, 54, was arrested for selling drugs and weapons, prosecutors say. Authorities allege Moschetto hired accomplices to burn down the practice of former associate.
Beam $k = 5$	Dr. Anthony Moschetto faces criminal solicitation, conspiracy, burglary, arson and weapons charges. "None of anything in this case has any evidentiary value," his attorney says.
Nucleus $\eta = 0.95$	Dr. Anthony Moschetto, 54, pleaded not guilty to charges Wednesday. Two men – identified as James Chmela and James Kalamaras – were named as accomplices.
Top-k $k = 30$	Dr. Anthony Moschetto is accused of providing police with weapons and prescription drugs. Authorities say he was part of a conspiracy to harm or kill a rival doctor. His attorney calls the allegations against his client "completely unsubstantiated"
Typical $\tau = 0.95$	Dr. Anthony Moschetto is charged with crimes including arson, conspiracy, burglary, prescription sale, weapons charges. His attorney says "none of anything in this case has any evidentiary value"

Story Generation (WritingPrompts)

Prompt	A kid doodling in a math class accidentally creates the world's first functional magic circle in centuries.
Reference	It was dark and Levi was pretty sure he was lying on his back. There was firelight flickering off of what was left of a ceiling. He could hear something but it was muffled. He was almost positive it was screaming. When he tried to move he felt an excruciating pain in his left side that caused him to cry out...
Nucleus $\eta = 0.95$	"Why do you draw a circle?" He asked as he handed the piece of paper over to her. She rolled her eyes and looked down at the picture, "Because I like math!" "Then what?" "I like to draw the circles." She ran a hand through her curls and looked away. "See, I have three circles." "Hmph, I didn't think I saw that in your class." A younger student brought up the ceiling and murmured...
Top-k $k = 30$	"Aha. Aah! That makes three." Professor Gosh sighed. "Oh my God, you had to have been there." "Well, I'm not sure if that's the right term, but this circle is going to work." "But I just told you that it wasn't working, didn't you, you little sh**? I'm not kidding! That circle is gon na work!" "So what now?"...
Temp $\tau = 1.0$	It was teachers duty to ensure that our kingdom is pure and pure and successful, however it does not mean we do not try to be as rebellious as any other kingdom. Most of our teachers are suitable for rulership, being knowledgeable in sadistic rec thunking of time stopping magic circle created by several apprentices...
Mirostat $\tau = 3$	ready to go home, when a little kid starts to sketch a circle. "No. Just stop. It's not working." "But the circle turns to smoke and then to a flame. It's not working, I know it is!" Another kid, who was also drawing a circle with his phone, starts to scream. The teacher, looking at the circle, lost his patience and started to yell...
Typical $\tau = 0.2$	As soon as he got the clock on his desk he jumped up. The noise it was making had changed his day forever. His hair, normally unkempt, now glistened and he had to have more hair because of it. "Yes!" He had declared triumphantly as he watched his father fumble around in the backpack in search of a little compass he always kept with him. He took it from the desk and quickly walked out the door to see what had happened. He wasn't the first child in the world to draw something...

Conclusion

We learned about several ways of decoding:

- Greedy decoding
- Beam search decoding
- Top k sampling
- Nucleus sampling
- Locally typical sampling

We also saw the effect of temperature on our decoding

References

- Neural Machine Translation with Reconstruction, Tu et al., 2017
- Six Challenges for Neural Machine Translation, Koehn et al., 2017
- The Curious Case of Neural Text Degeneration, Holtzman et al., 2020
- Locally Typical Sampling, Meister et al., 2022
- Beam Decoding with Controlled Patience, Kasai et al., 2022