Decoding from language models

CS685 Fall 2021

Advanced Natural Language Processing

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

Review: sequence-to-sequence models

- we'll use French (*f*) to English (*e*) as a running example
- **goal**: given French sentence *f* with tokens f_1, f_2, … f_n produce English translation *e* with tokens e_1, e_2, … e_m
- **real goal**: compute $\arg\max_e p(e \mid f)$

Review: sequence-to-sequence models

- let's use an NN to directly model $p(e \mid f)$

$$
\begin{aligned}
p(e \mid f) &= p(e_1, e_2, \ldots, e_l \mid f) \\
&= p(e_1 \mid f) \cdot p(e_2 \mid e_1, f) \cdot p(e_3 \mid e_2, e_1, f) \cdot \ldots \\
&= \prod_{i=1}^{L} p(e_i \mid e_1, \ldots, e_{i-1}, f)
\end{aligned}
$$

seq2seq models

- use two different NNs to model $\prod_{i=1}^{L} p(e_i \,|\, e_1, \ldots, e_{i-1}, f)$

- first we have the *encoder*, which encodes the French sentence *f*
- then, we have the *decoder,* which produces the English sentence *e*

We’ve already talked about training these models… what about test-time usage?

decoding

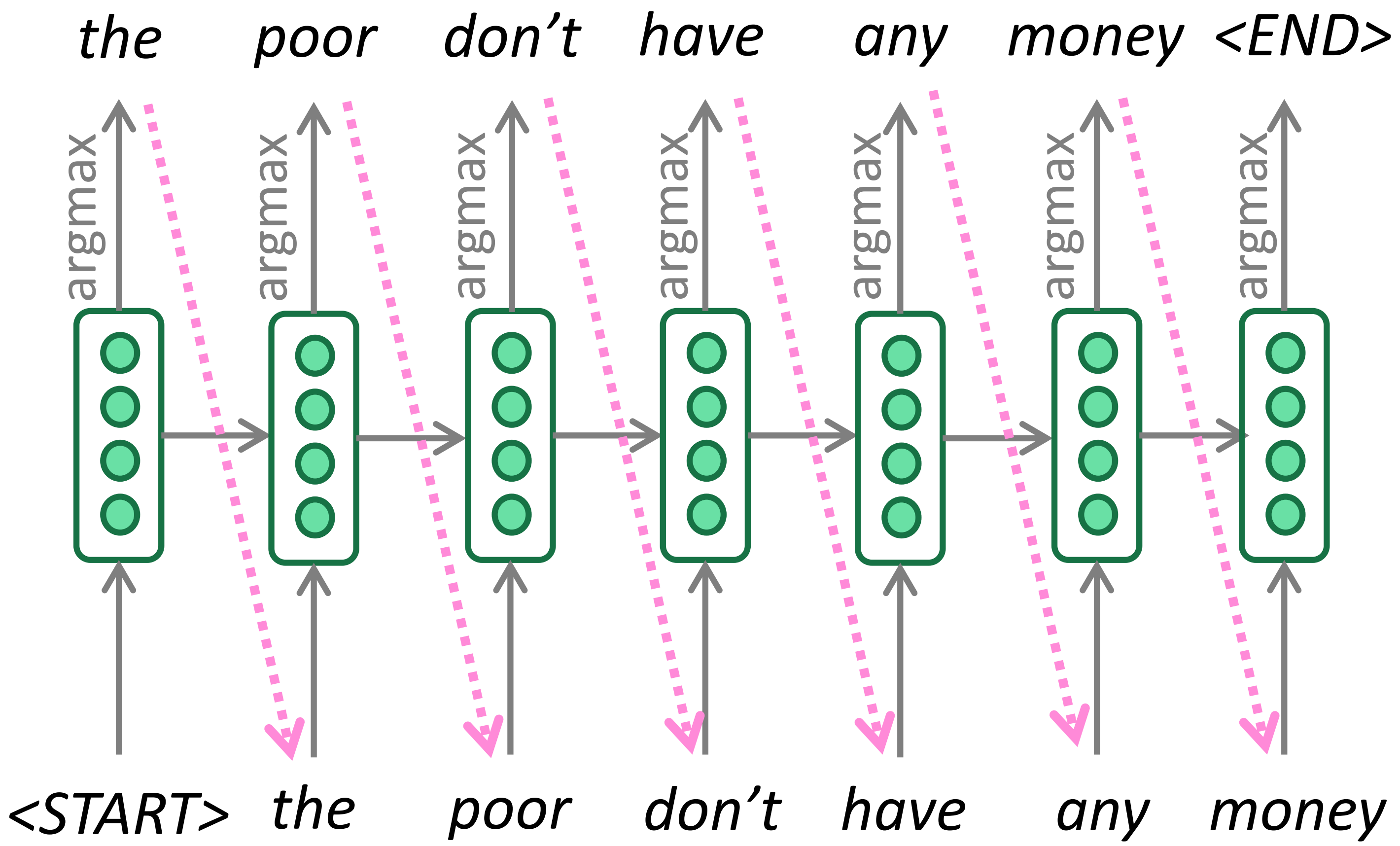
- given that we trained a seq2seq model, how do we find the most probable English sentence?
- more concretely, how do we find

$$\arg\max \prod_{i=1}^{L} p(e_i \mid e_1, \ldots, e_{i-1}, f)$$

- can we enumerate all possible English sentences *e*?

decoding

- given that we trained a seq2seq model, how do we find the most probable English sentence?
- easiest option: **greedy decoding**



issues?

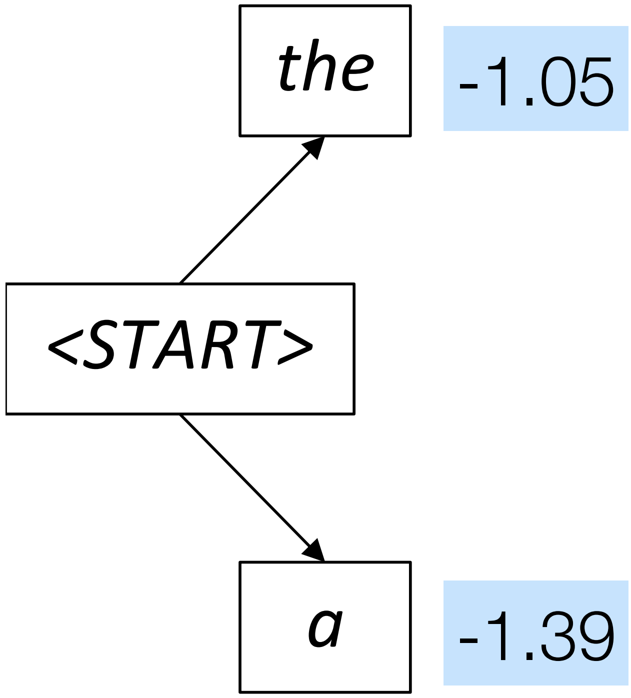
Beam search

- in greedy decoding, we cannot go back and revise previous decisions!
  - *les pauvres sont démunis (the poor don't have any money)*
  - → *the* ____
  - → *the poor* ____
  - → *the poor are* ____
- fundamental idea of beam search: explore several different hypotheses instead of just a single one
  - keep track of *k* most probable partial translations at each decoder step instead of just one!

    the beam size *k* is usually 5-10

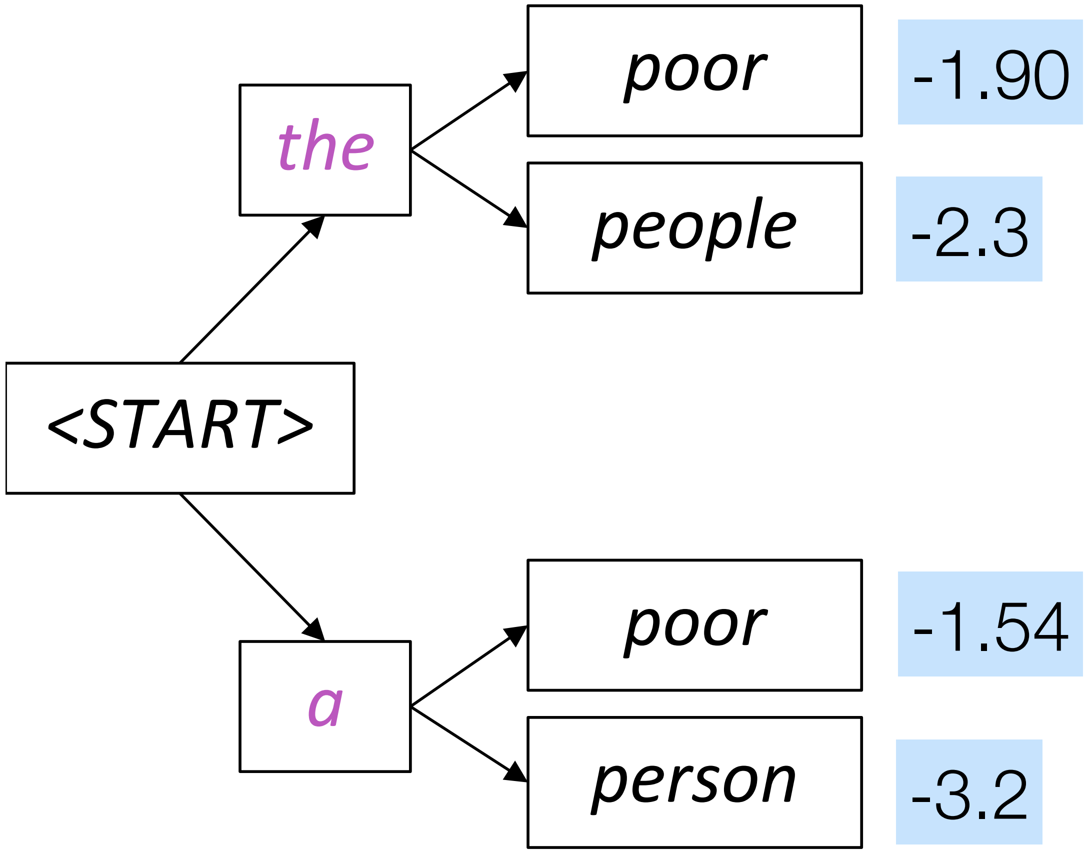
Beam search decoding: example

Beam size = 2



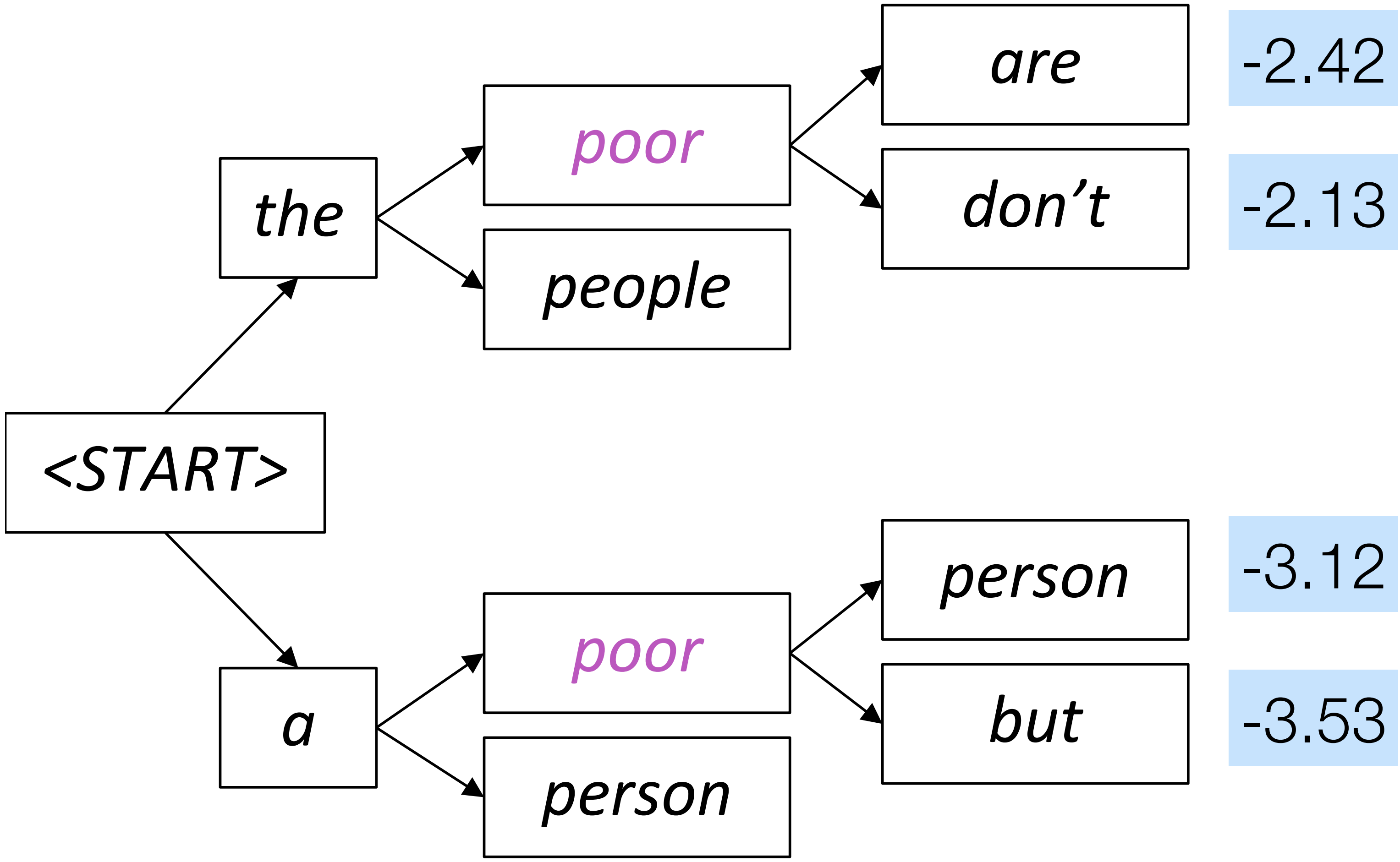
Beam search decoding: example

Beam size = 2



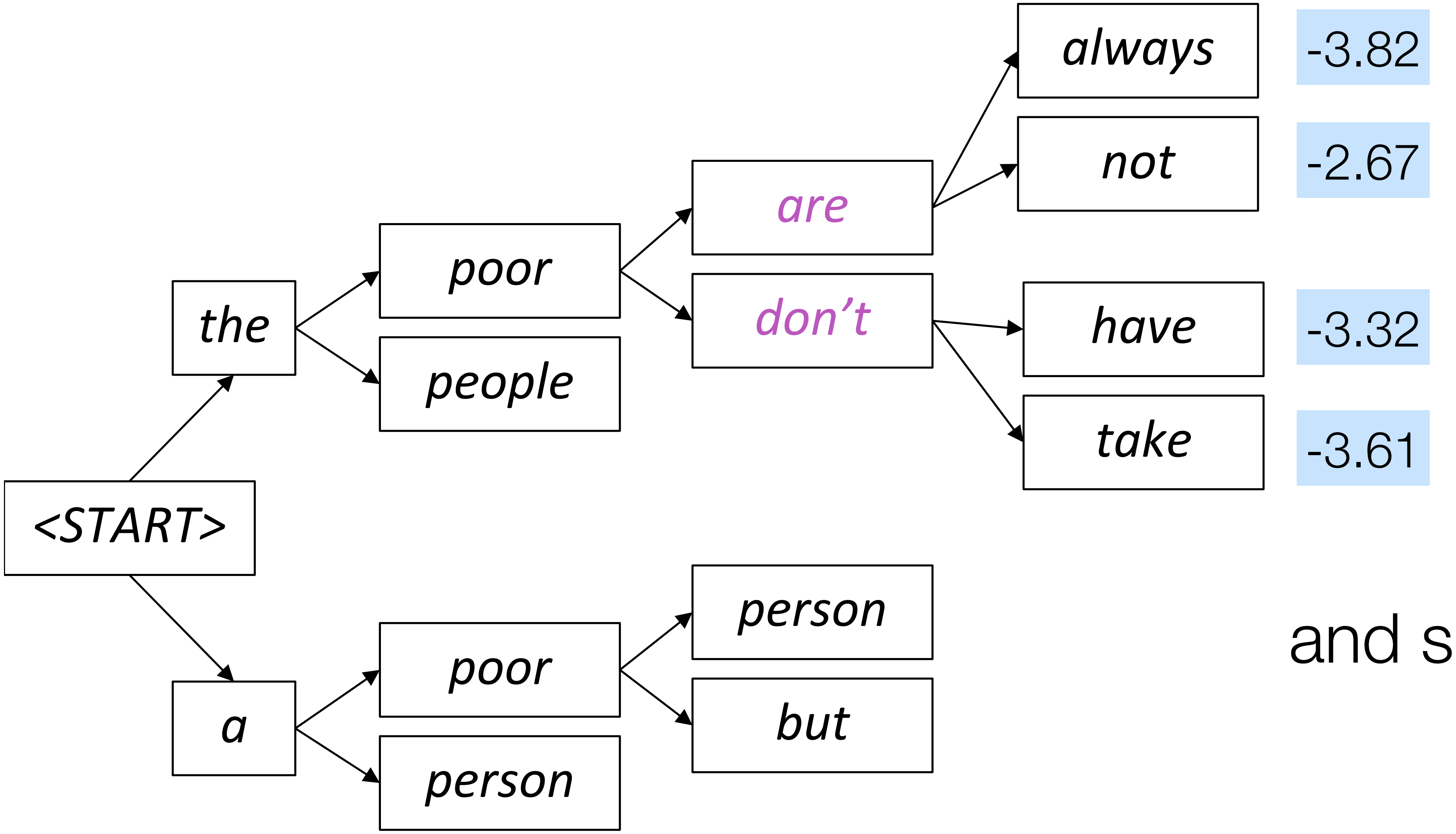
Beam search decoding: example

Beam size = 2



Beam search decoding: example

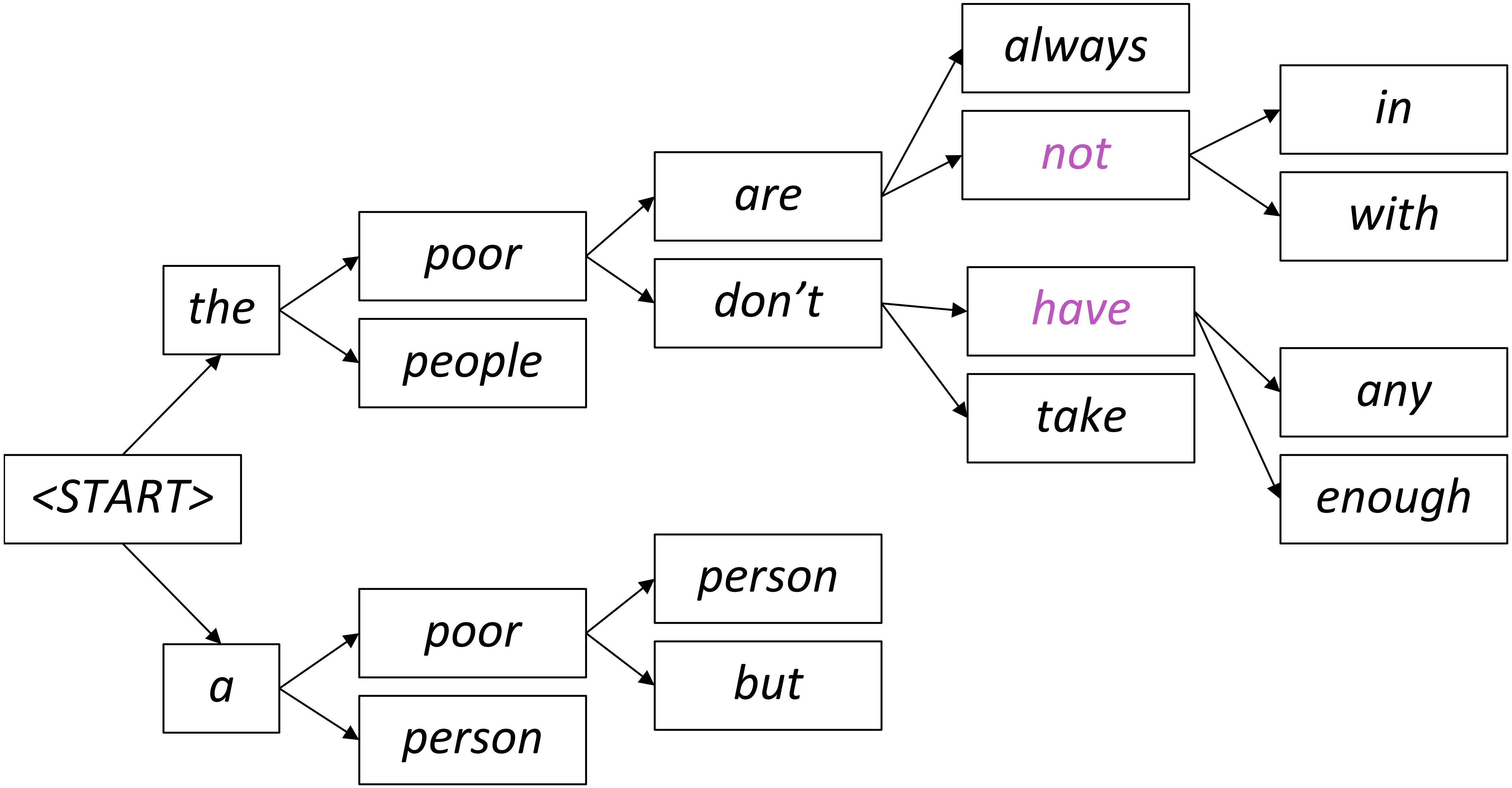
Beam size = 2



and so on…

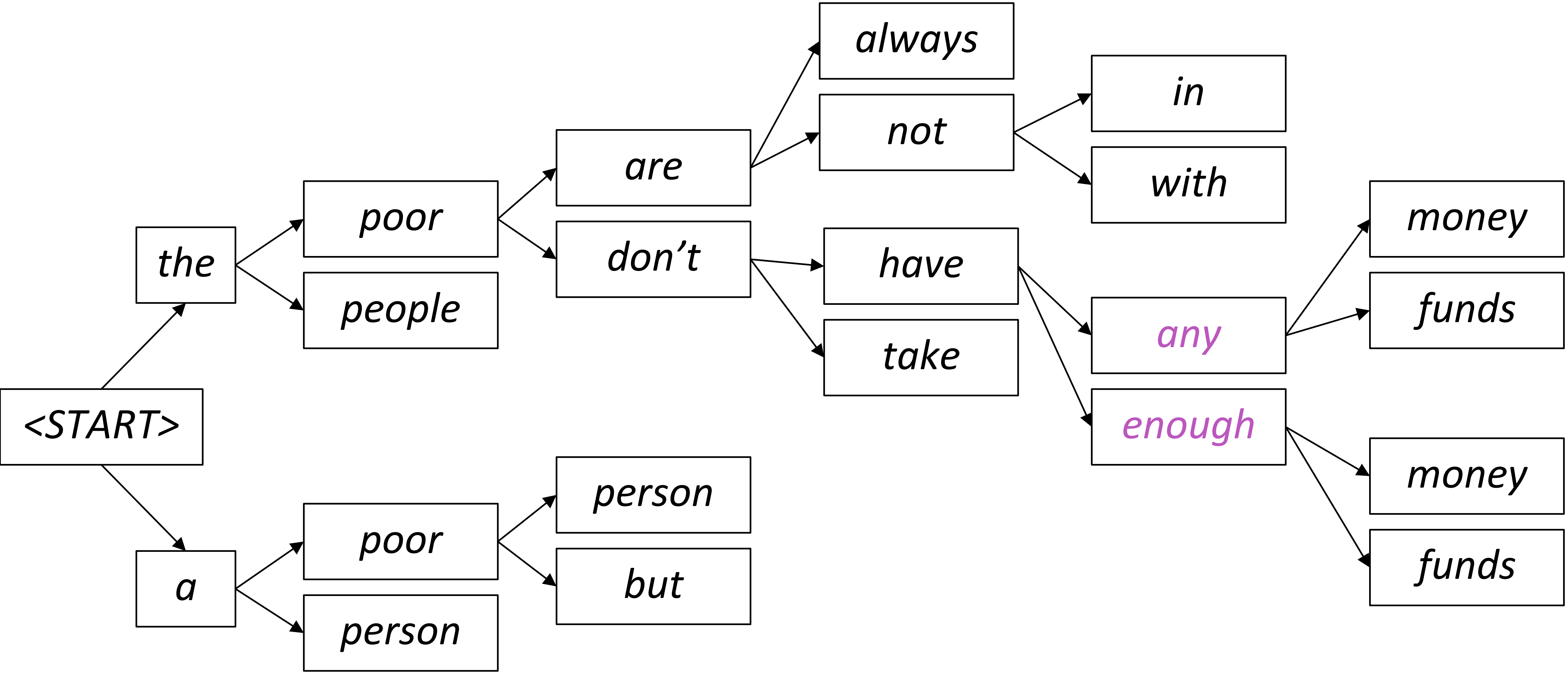
Beam search decoding: example

Beam size = 2



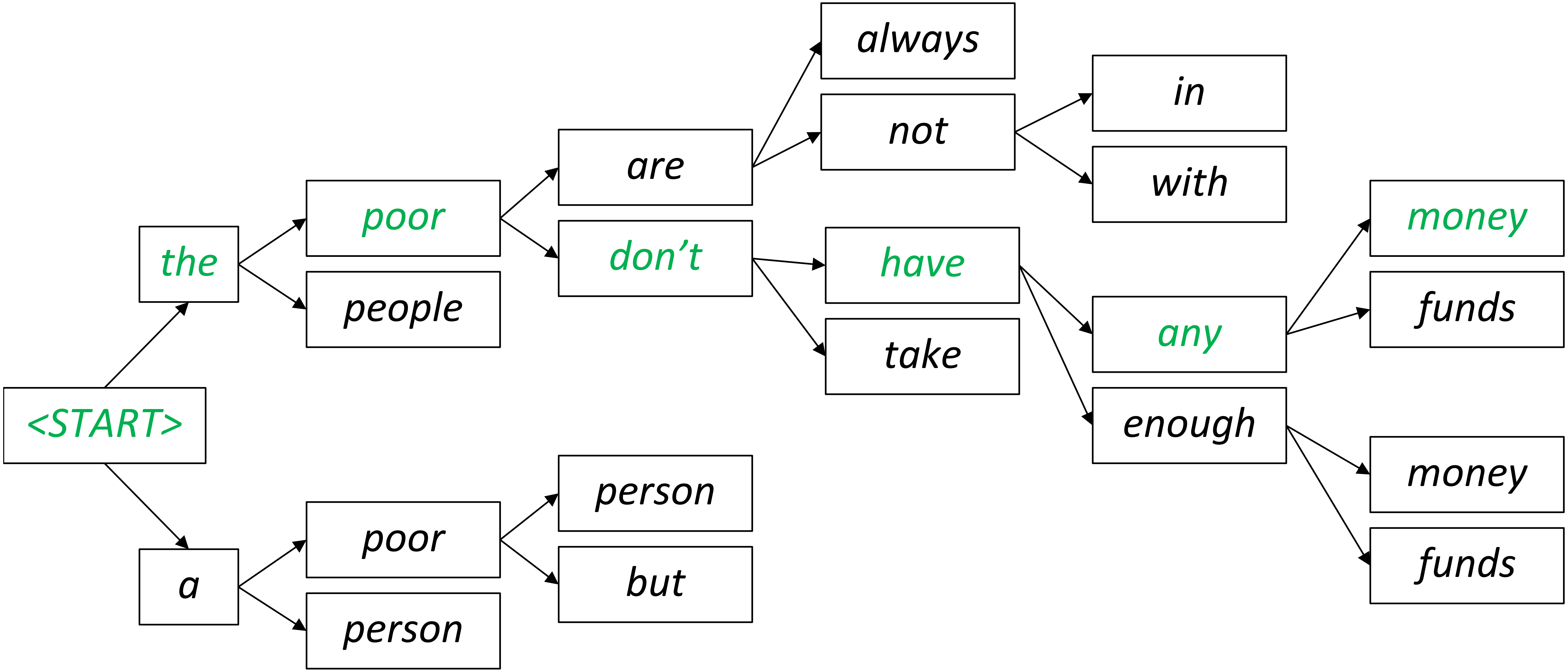
Beam search decoding: example

Beam size = 2



Beam search decoding: example

Beam size = 2



does beam search always produce the *best* translation (i.e., does it always find the argmax?)

what are the termination conditions for beam search?

What if we want to maximize output *diversity* rather than find a highly probable sequence?

What’s the effect of changing beam size *k*?

- Small *k* has similar problems to greedy decoding (*k*=1)
  - Ungrammatical, unnatural, nonsensical, incorrect

- Larger *k* means you consider more hypotheses
  - Increasing *k* reduces some of the problems above
  - Larger *k* is more computationally expensive
  - But increasing *k* can introduce other problems:
    - For NMT, increasing *k* too much decreases BLEU score (Tu et al, Koehn et al). This is primarily because large-k beam search produces too-short translations (even with score normalization!)
    - In open-ended tasks like chit-chat dialogue, large *k* can make output more generic (see next slide)

Neural Machine Translation with Reconstruction, Tu et al, 2017 https://arxiv.org/pdf/1611.01874.pdf
Six Challenges for Neural Machine Translation, Koehn et al, 2017 https://arxiv.org/pdf/1706.03872.pdf

Effect of beam size in chitchat dialogue

I mostly eat a fresh and raw diet, so I save on groceries

Human chit-chat partner

Beam size	Model response
1	*I love to eat healthy and eat healthy*
2	*That is a good thing to have*
3	*I am a nurse so I do not eat raw food*
4	*I am a nurse so I am a nurse*
5	*Do you have any hobbies?*
6	*What do you do for a living?*
7	*What do you do for a living?*
8	*What do you do for a living?*

Low beam size:
More on-topic but nonsensical;
bad English

High beam size:
Converges to safe, "correct" response, but it's generic and less relevant

Sampling-based decoding

Both of these are **more efficient** than beam search – no multiple hypotheses

- Pure sampling
  - On each step *t*, randomly sample from the probability distribution P_t to obtain your next word.
  - Like greedy decoding, but sample instead of argmax.

- Top-n sampling*
  - On each step *t*, randomly sample from $P_{t,}$ restricted to just the top-n most probable words
  - Like pure sampling, but truncate the probability distribution
  - *n=1* is greedy search, *n=V* is pure sampling
  - Increase *n* to get more diverse/risky output
  - Decrease *n* to get more generic/safe output

*Usually called top-*k* sampling, but here we're avoiding confusion with beam size *k*

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.

The Curious Case of Neural Text Degeneration, Holtzman et al., 2020

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

WebText

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.

The Curious Case of Neural Text Degeneration, 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.

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.

The Curious Case of Neural Text Degeneration, 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.

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.

The Curious Case of Neural Text Degeneration, Holtzman et al., 2020

Decoding algorithms: in summary

- **Greedy decoding** is a simple method; gives low quality output
- **Beam search** (especially with high beam size) searches for high-probability output
  - Delivers better quality than greedy, but if beam size is too high, can return high-probability but unsuitable output (e.g. generic, short)
- **Sampling methods** are a way to get more diversity and randomness
  - Good for open-ended / creative generation (poetry, stories)
  - Top-n sampling allows you to control diversity