Text generation:

decoding / evaluation

CS 685, Fall 2020

Advanced Natural Language Processing

Mohit Iyyer

College of Information and Computer Sciences

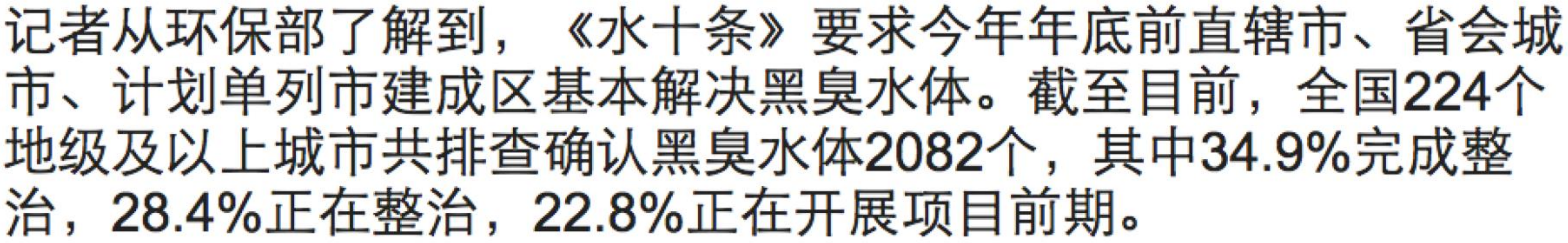
University of Massachusetts Amherst

some slides adapted from Marine Carpuat, Richard Socher, & Abigail See

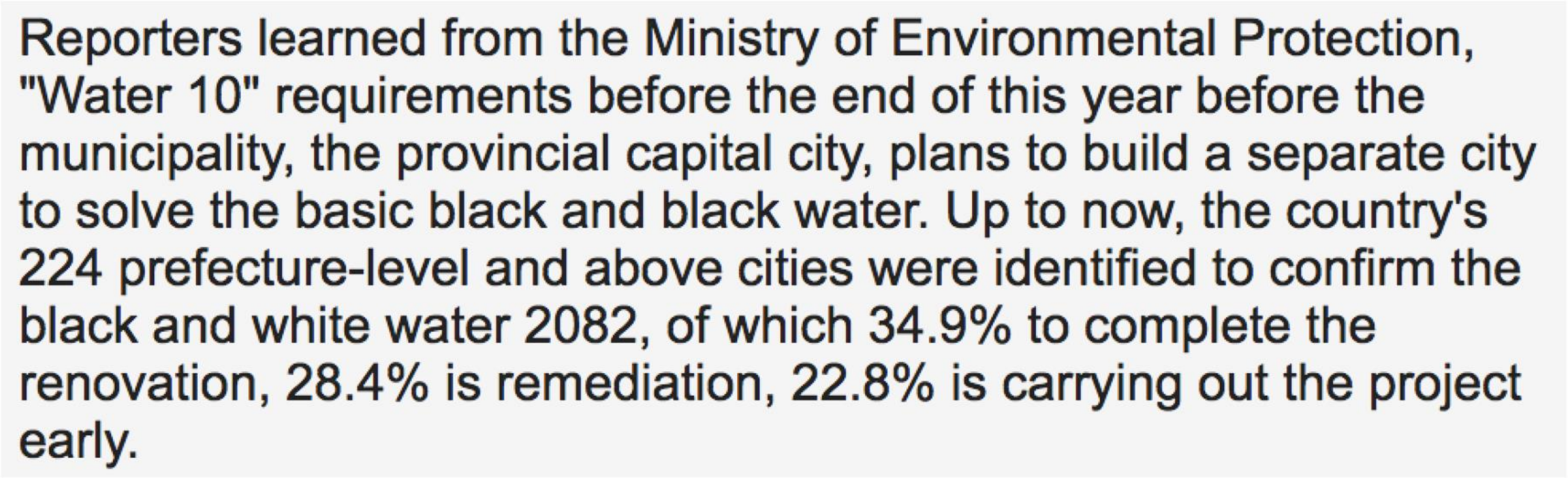
stuff from last time…

- More implementation classes?

How Good is Machine Translation? Chinese > English

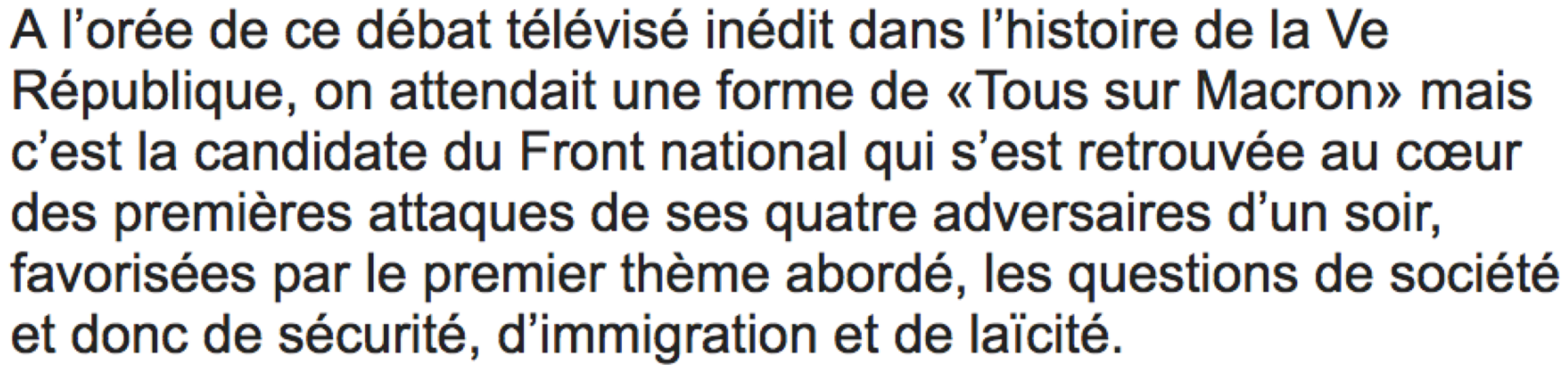


记者从环保部了解到，《水十条》要求今年年底前直辖市、省会城市、计划单列市建成区基本解决黑臭水体。截至目前，全国224个地级及以上城市共排查确认黑臭水体2082个，其中34.9%完成整治，28.4%正在整治，22.8%正在开展项目前期。

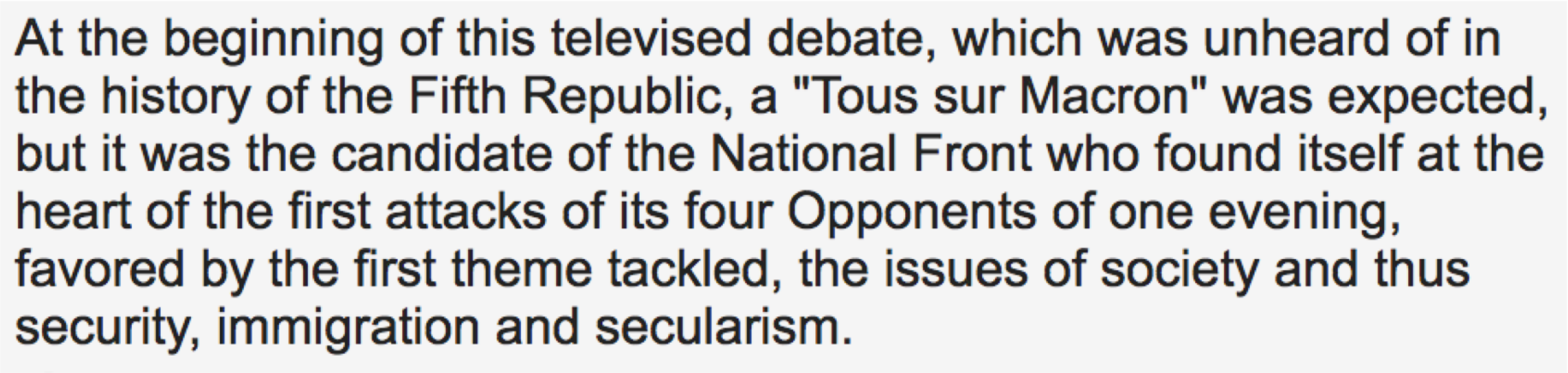


Reporters learned from the Ministry of Environmental Protection, "Water 10" requirements before the end of this year before the municipality, the provincial capital city, plans to build a separate city to solve the basic black and black water. Up to now, the country's 224 prefecture-level and above cities were identified to confirm the black and white water 2082, of which 34.9% to complete the renovation, 28.4% is remediation, 22.8% is carrying out the project early.

How Good is Machine Translation?
French > English



A l’orée de ce débat télévisé inédit dans l’histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c’est la candidate du Front national qui s’est retrouvée au cœur des premières attaques de ses quatre adversaires d’un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d’immigration et de laïcité.



At the beginning of this televised debate, which was unheard of in the history of the Fifth Republic, a "Tous sur Macron" was expected, but it was the candidate of the National Front who found itself at the heart of the first attacks of its four Opponents of one evening, favored by the first theme tackled, the issues of society and thus security, immigration and secularism.

What is MT good (enough) for?

- **Assimilation:** reader initiates translation, wants to know content
  - User is tolerant of inferior quality
  - Focus of majority of research
- **Communication:** participants in conversation don’t speak same language
  - Users can ask questions when something is unclear
  - Chat room translations, hand-held devices
  - Often combined with speech recognition
- **Dissemination:** publisher wants to make content available in other languages
  - High quality required
  - Almost exclusively done by human translators

review: neural MT

- 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: neural MT

- 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

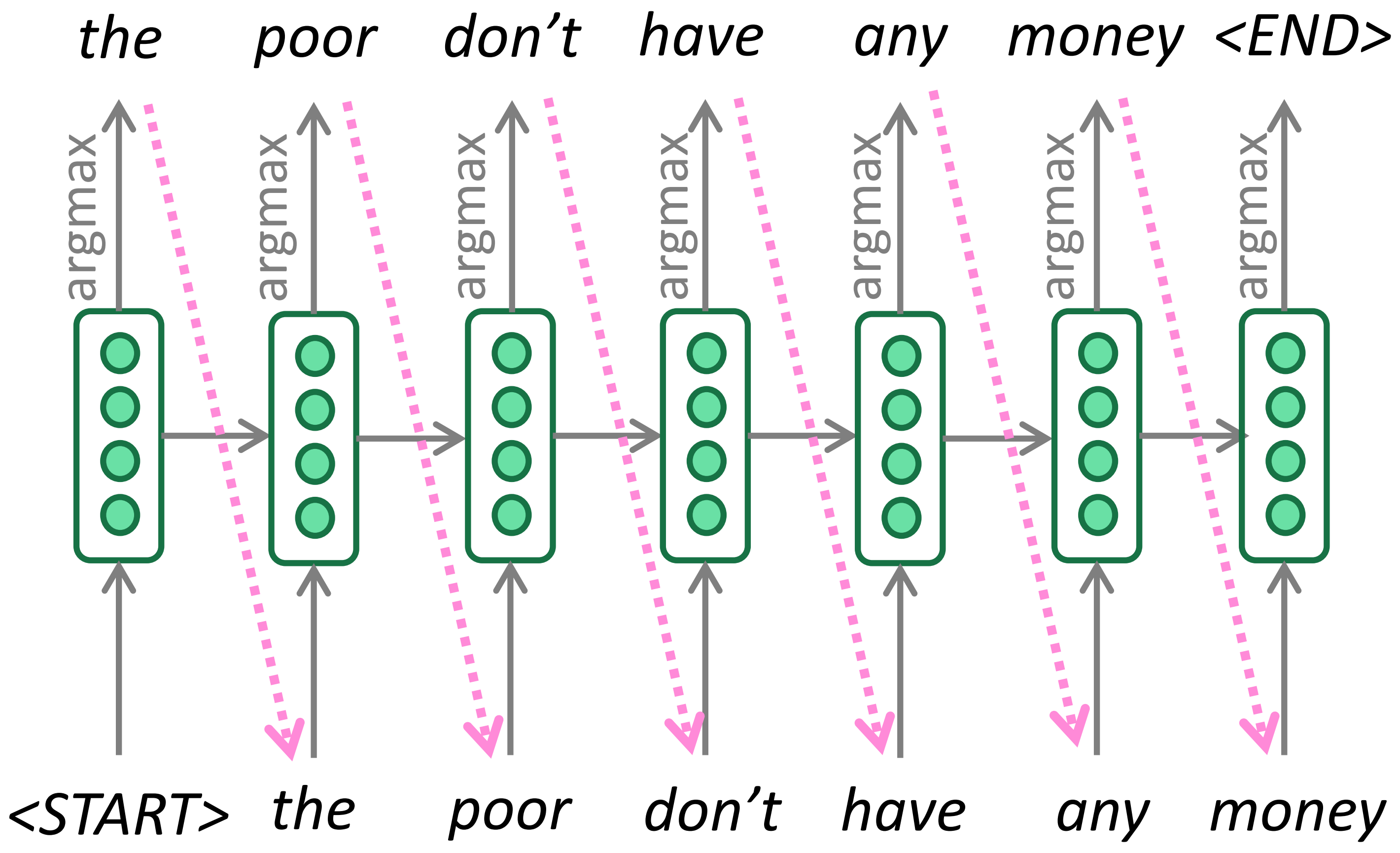
- given that we trained a seq2seq model, how do we find the most probable English sentence?
- more concretely, how do we find

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

- 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

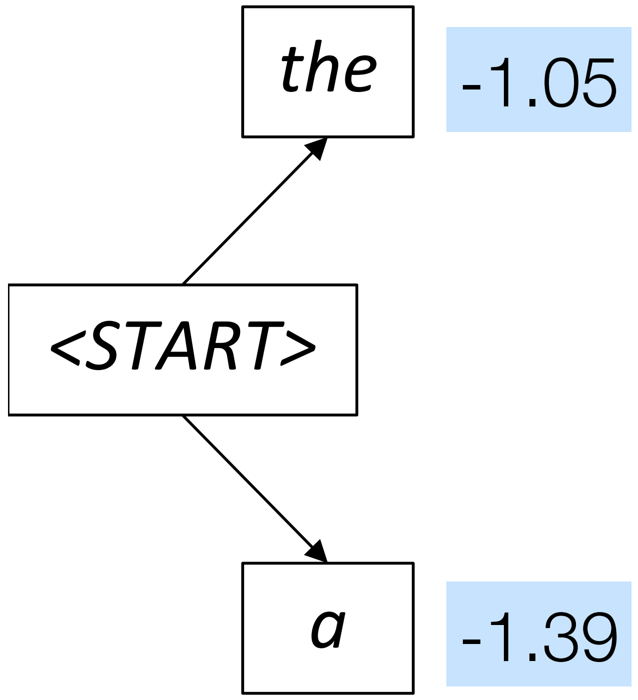
- in greedy decoding, we cannot go back and revise previous decisions!
  - *les pauvres sont démunis (the poor don't have any money)*
  - → *the ____*
  - → *the poor ____*
  - → *the poor are ____*

- 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

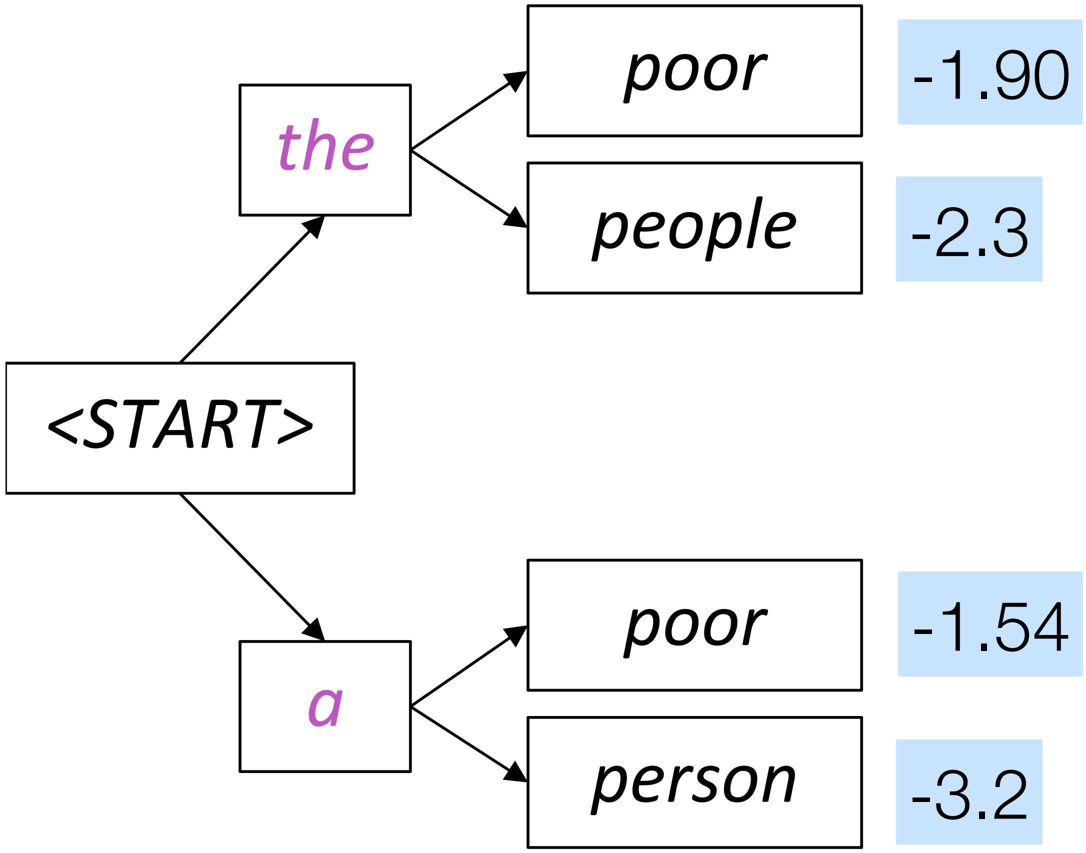
Beam search decoding: example

Beam size = 2



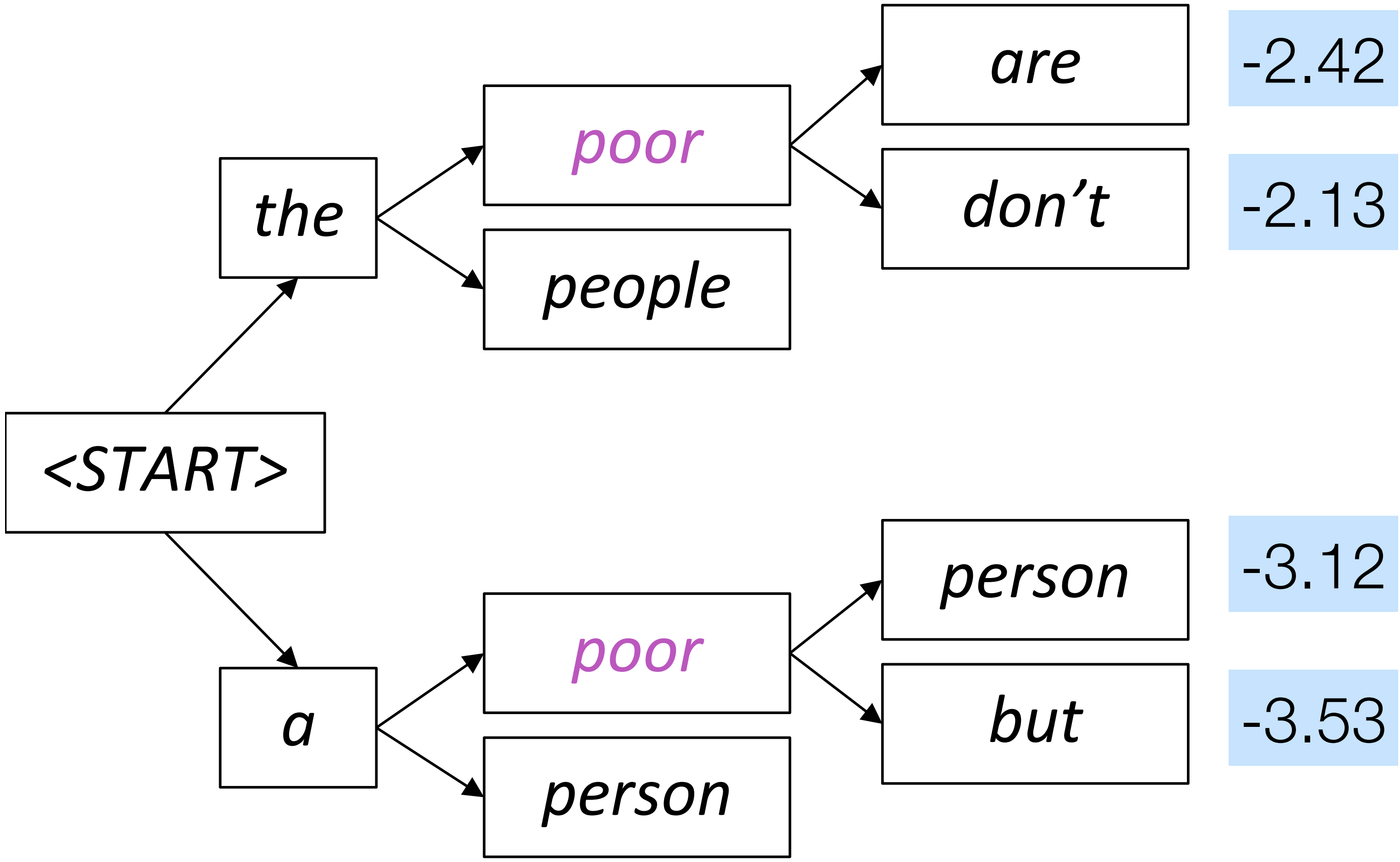
Beam search decoding: example

Beam size = 2



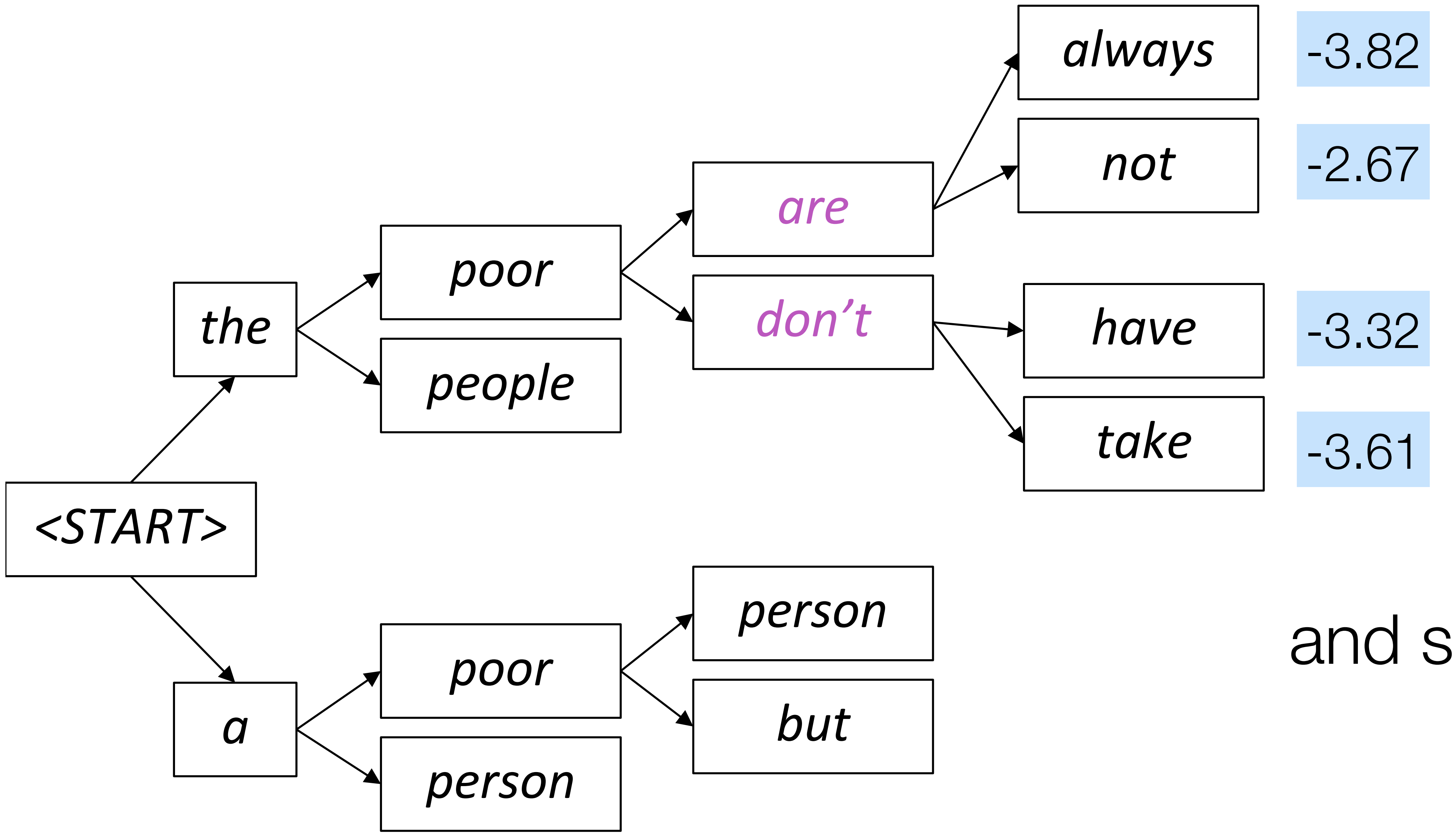
Beam search decoding: example

Beam size = 2



Beam search decoding: example

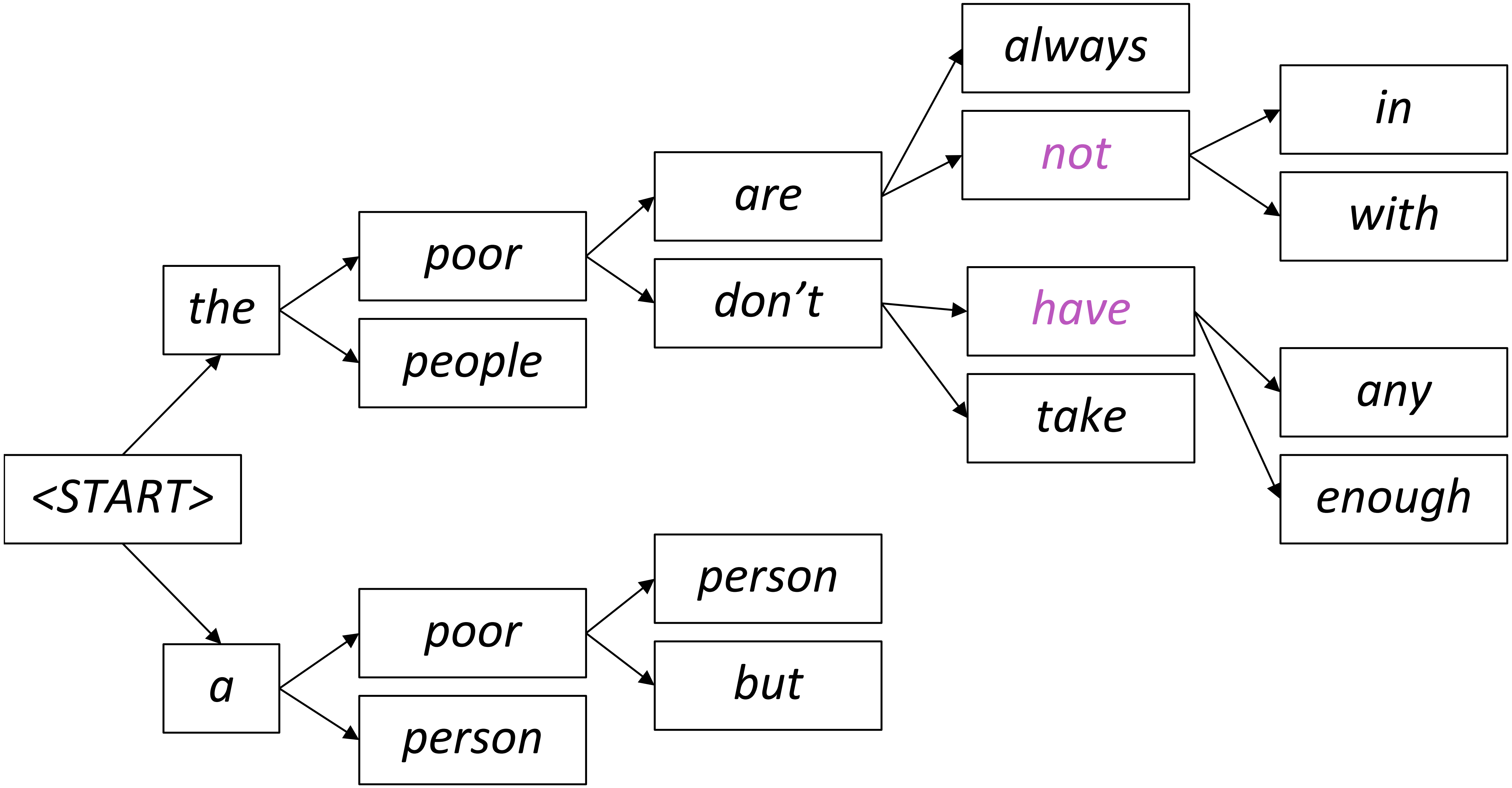
Beam size = 2



and so on…

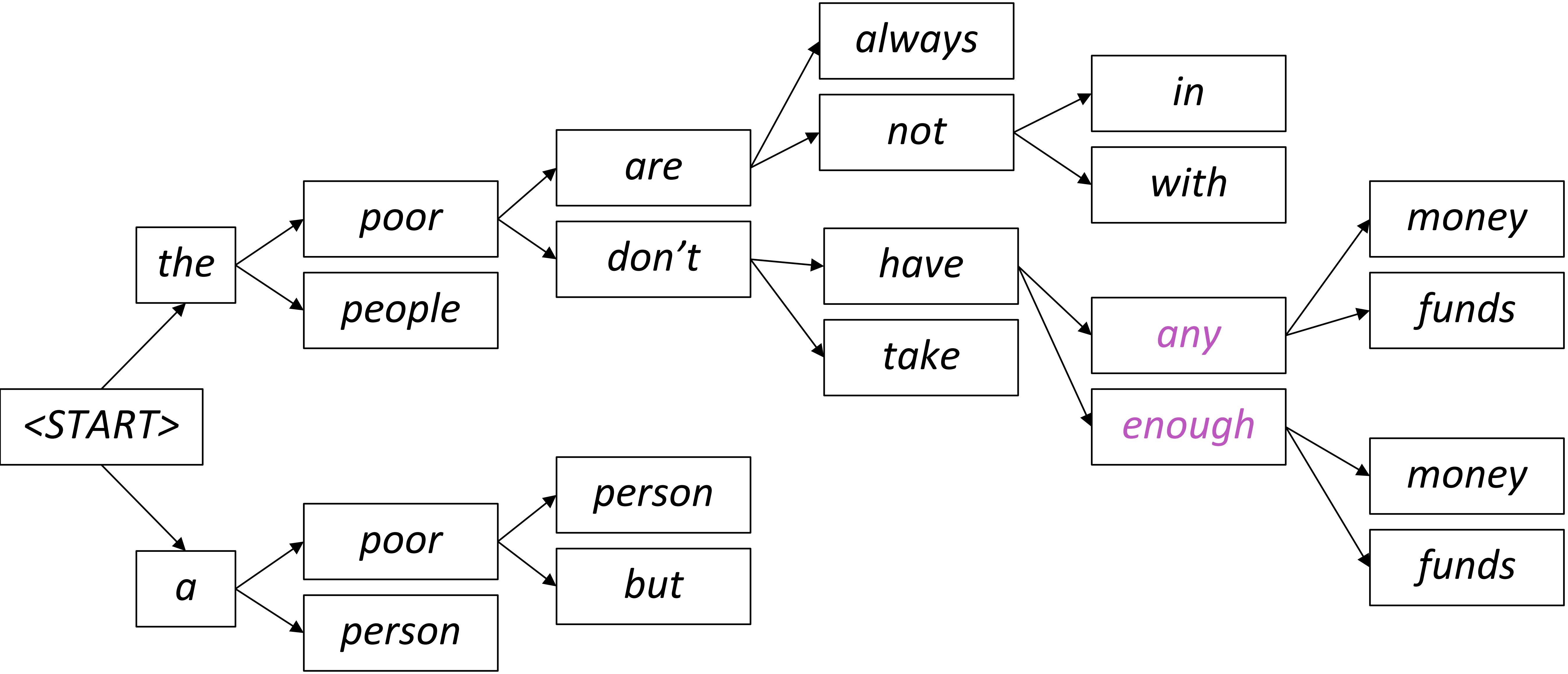
Beam search decoding: example

Beam size = 2



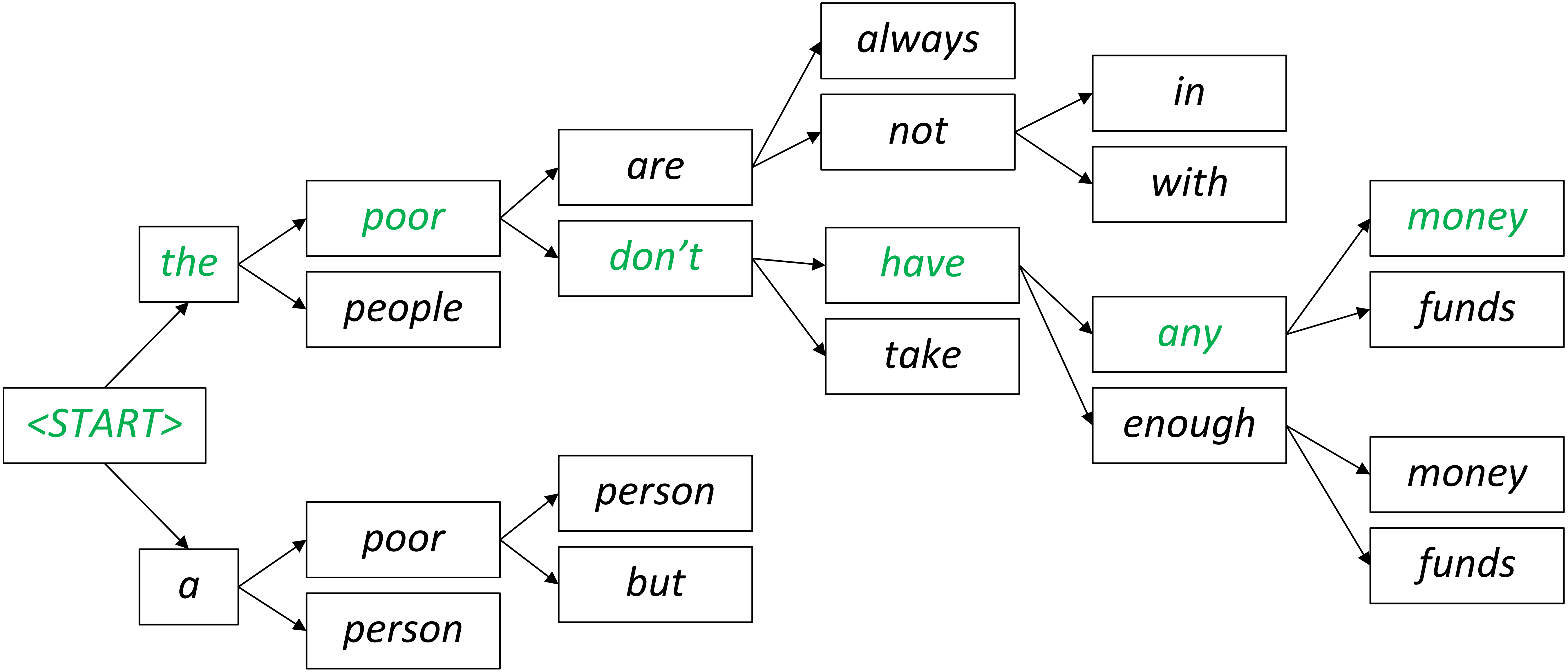
Beam search decoding: example

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

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.

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

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.

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

onto evaluation…

How good is a translation?
Problem: no single right answer

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Evaluation

- How good is a given machine translation system?

- Many different translations acceptable

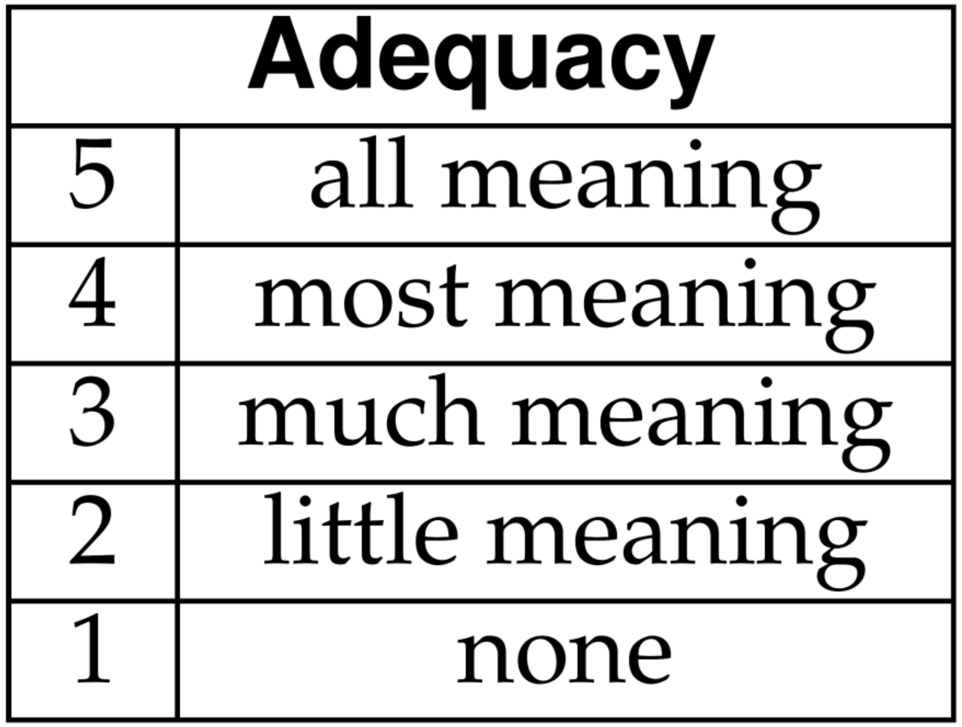
- Evaluation metrics
  - Subjective judgments by human evaluators
  - Automatic evaluation metrics
  - Task-based evaluation

Adequacy and Fluency

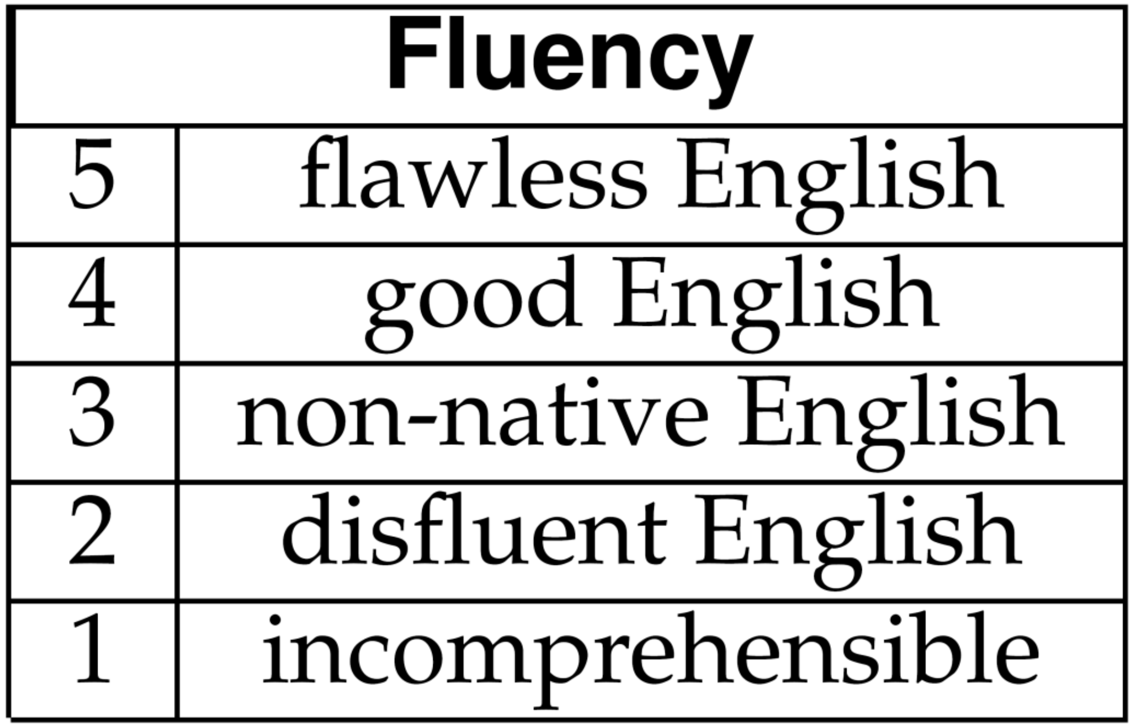
- Human judgment
  - Given: machine translation output
  - Given: input and/or reference translation
  - Task: assess quality of MT output

- Metrics
  - **Adequacy:** does the output convey the meaning of the input sentence? Is part of the message lost, added, or distorted?
  - **Fluency:** is the output fluent? Involves both grammatical correctness and idiomatic word choices.

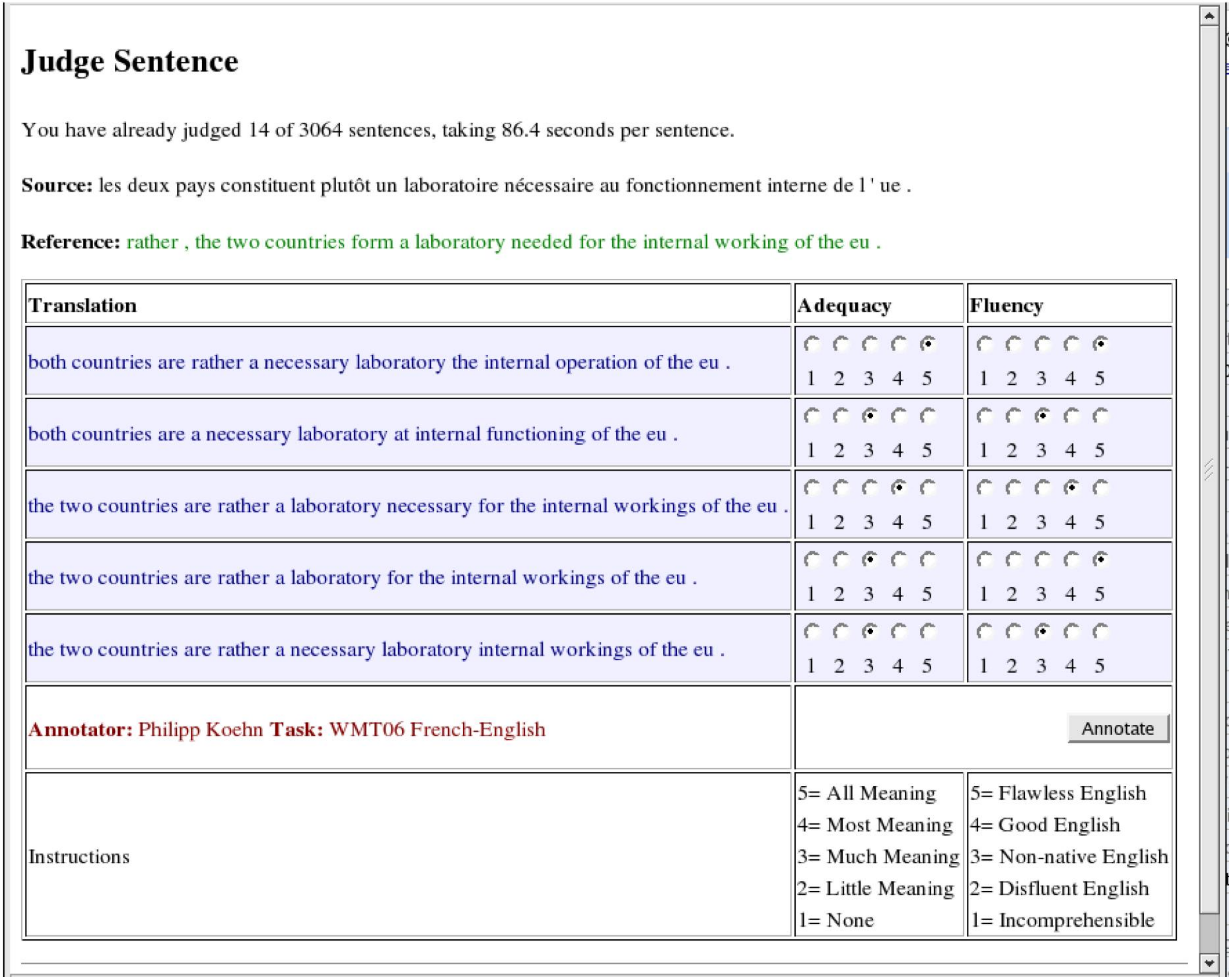
Fluency and Adequacy: Scales



Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none



Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	○ ○ ○ ○ ● 1 2 3 4 5	○ ○ ○ ○ ● 1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ● ○ ○ 1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	○ ○ ○ ● ○ 1 2 3 4 5	○ ○ ○ ● ○ 1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ○ ○ ● 1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ● ○ ○ 1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English	Annotate	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Let's try:
rate fluency & adequacy on 1-5 scale

- Source:
  N'y aurait-il pas comme une vague hypocrisie de votre part ?
- Reference:
  Is there not an element of hypocrisy on your part?
- System1:
  Would it not as a wave of hypocrisy on your part?
- System2:
  Is there would be no hypocrisy like a wave of your hand?
- System3:
  Is there not as a wave of hypocrisy from you?

what are some issues with human evaluation?

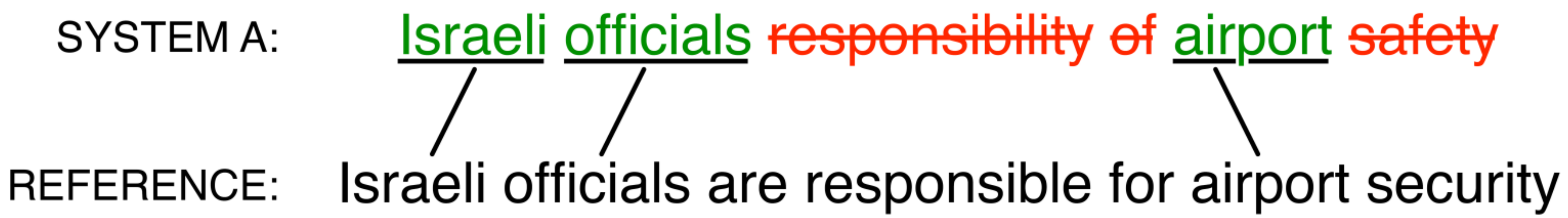
Automatic Evaluation Metrics

- Goal: computer program that computes quality of translations

- Advantages: low cost, optimizable, consistent

- Basic strategy
  - Given: MT output
  - Given: human reference translation
  - Task: compute similarity between them

Precision and Recall of Words



SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

Precision

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

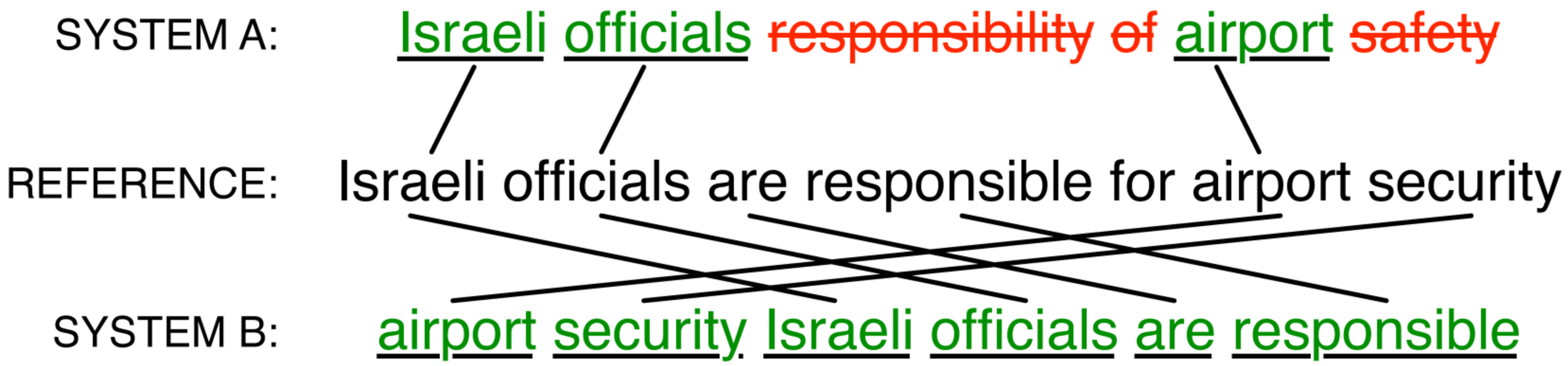
Recall

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

F-measure

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall of Words



SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

Metric	System A	System B
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

flaw: no penalty for reordering

BLEU
Bilingual Evaluation Understudy

N-gram overlap between machine translation output and reference translation

Compute precision for n-grams of size 1 to 4

Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \ (\prod_{i=1}^{4} \textit{precision}_i)^{\frac{1}{4}}$$

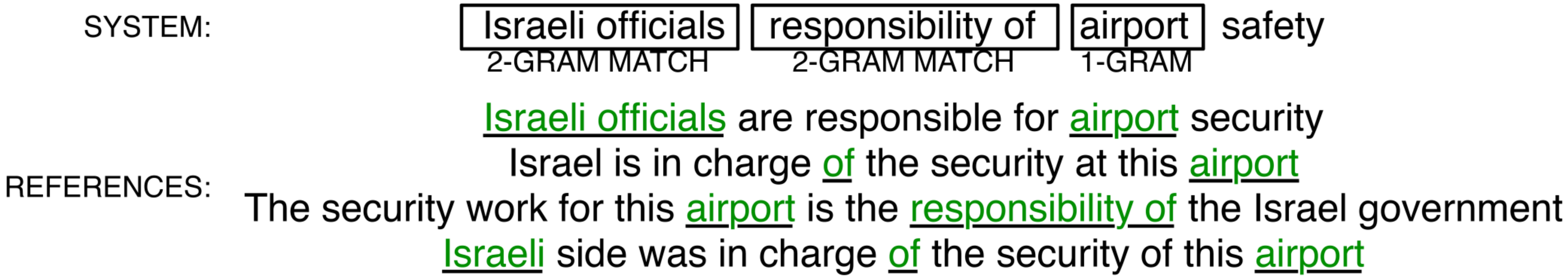
Typically computed over the entire corpus, not single sentences

Multiple Reference Translations

To account for variability, use multiple reference translations

- n-grams may match in any of the references
- closest reference length used

Example



SYSTEM: Israeli officials responsibility of airport safety

2-GRAM MATCH   2-GRAM MATCH   1-GRAM

REFERENCES:

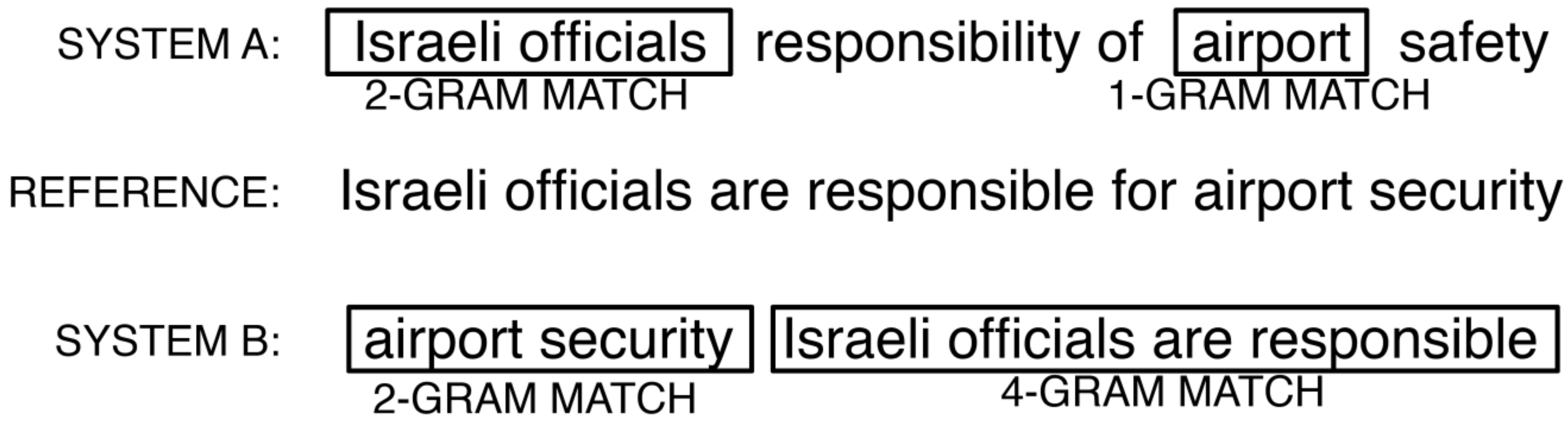
Israeli officials are responsible for airport security

Israel is in charge of the security at this airport

The security work for this airport is the responsibility of the Israel government

Israeli side was in charge of the security of this airport

BLEU examples



SYSTEM A: Israeli officials responsibility of airport safety

2-GRAM MATCH

1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

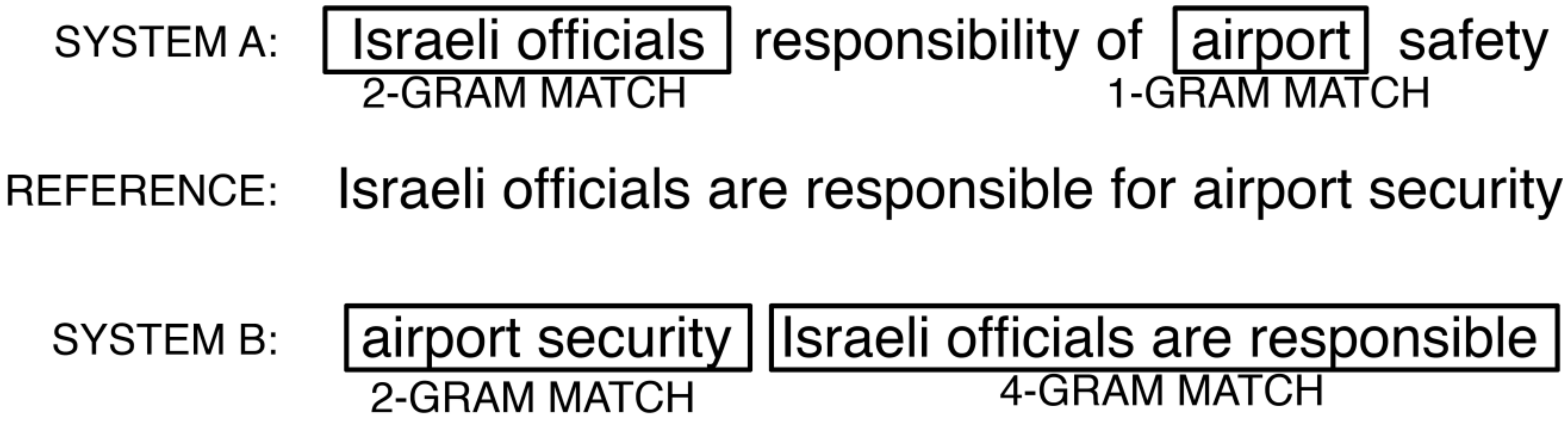
SYSTEM B: airport security Israeli officials are responsible

2-GRAM MATCH

4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

BLEU examples



SYSTEM A: Israeli officials responsibility of airport safety

2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

2-GRAM MATCH 4-GRAM MATCH

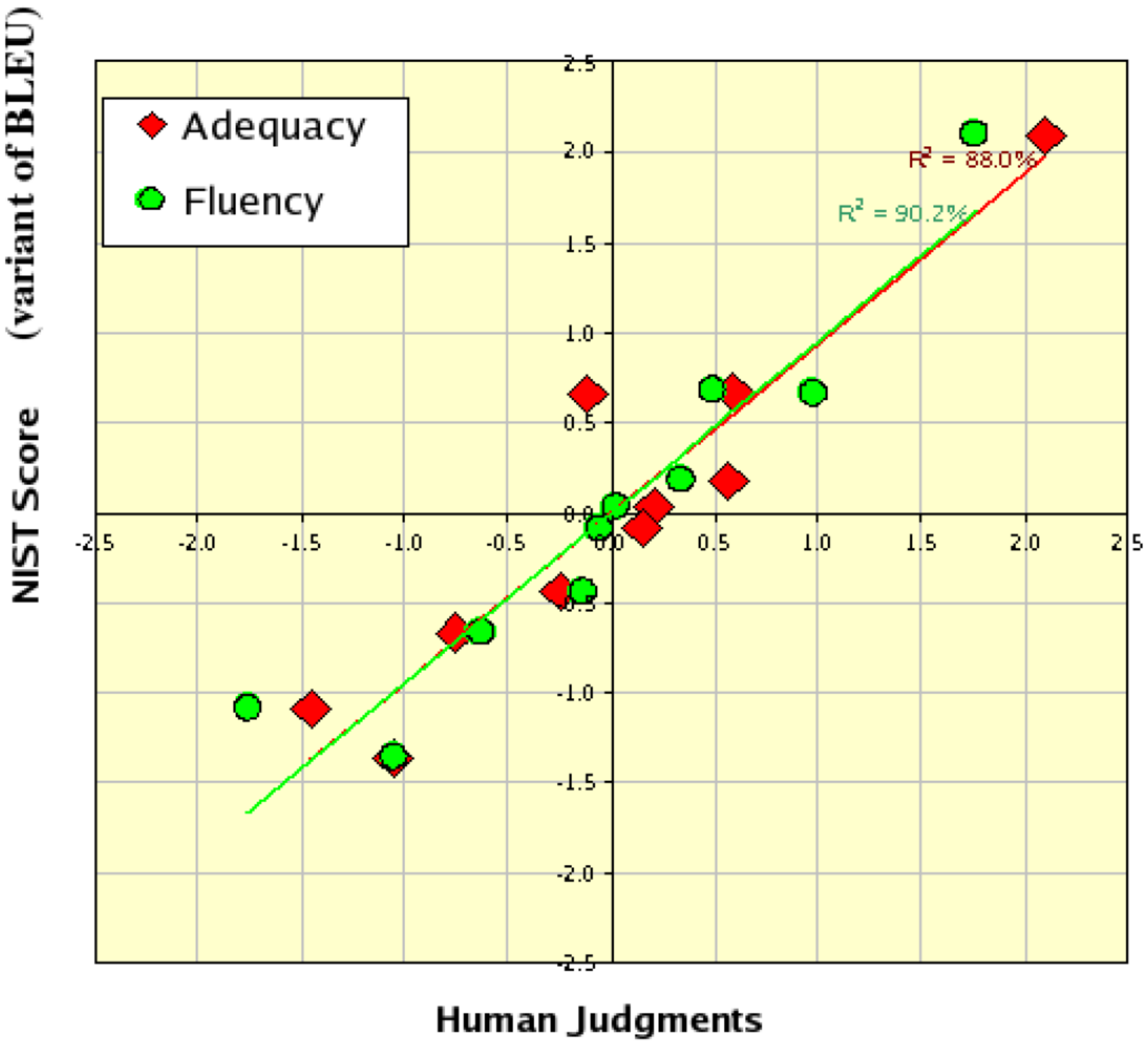
why does BLEU not account for recall?

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

what are some drawbacks of BLEU?

- all words/n-grams treated as equally relevant
- operates on local level
- scores are meaningless (absolute value not informative)
- human translators also score low on BLEU

Yet automatic metrics such as BLEU correlate with human judgement



Can we include *learned* components in our evaluation metrics?

BLEURT (BLEU + BERT)

- Take a pretrained BERT, and fine-tune it on a variety of synthetic tasks with perturbed data
  - Synthetic data involves a sentence ***z*** and "perturbed" version ***z'***
  - Objectives include many regression tasks (e.g., predict BLEU, ROUGE, backtranslation likelihood)
- Then, fine-tune the resulting model on small supervised datasets of human quality judgments

BLEURT (BLEU + BERT)

- Take a pretrained BERT, and fine-tune it on a variety of synthetic tasks with perturbed data
  - Synthetic data involves a sentence ***z*** and "perturbed" version ***z'***
  - Objectives include many regression tasks (e.g., predict BLEU, ROUGE, backtranslation likelihood)
- Then, fine-tune the resulting model on small supervised datasets of human quality judgments

Higher correlation with human judgments than just BLEU, but has limitations…