

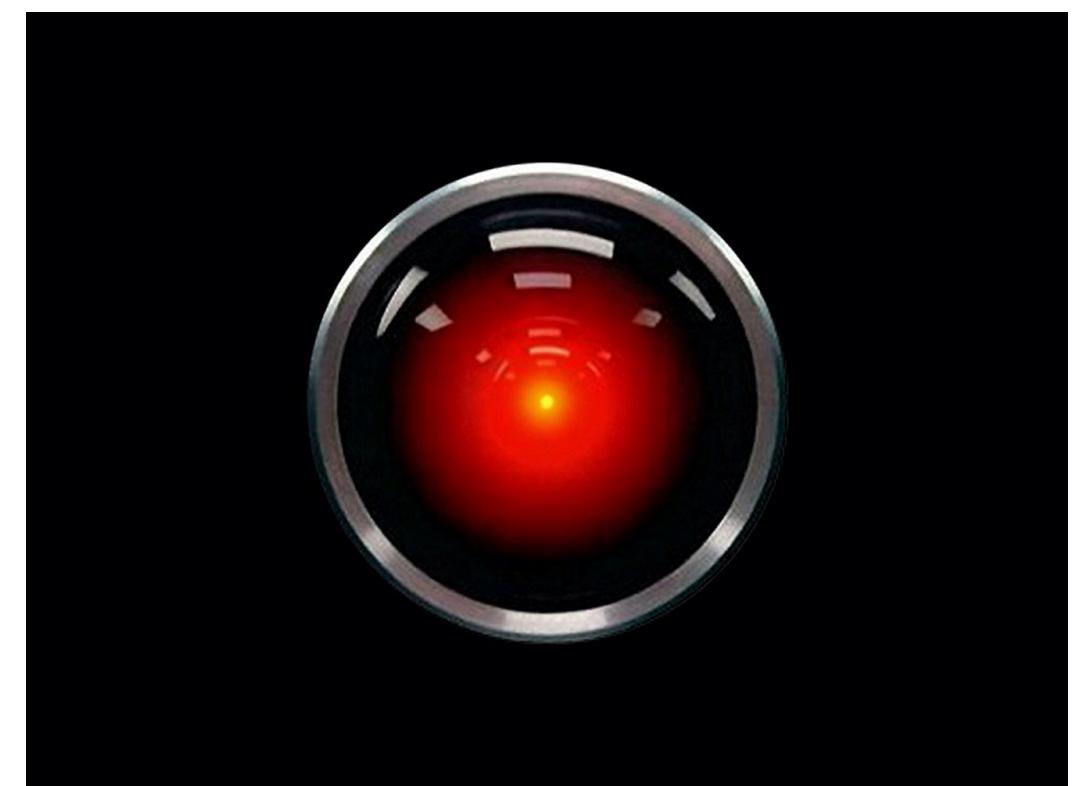
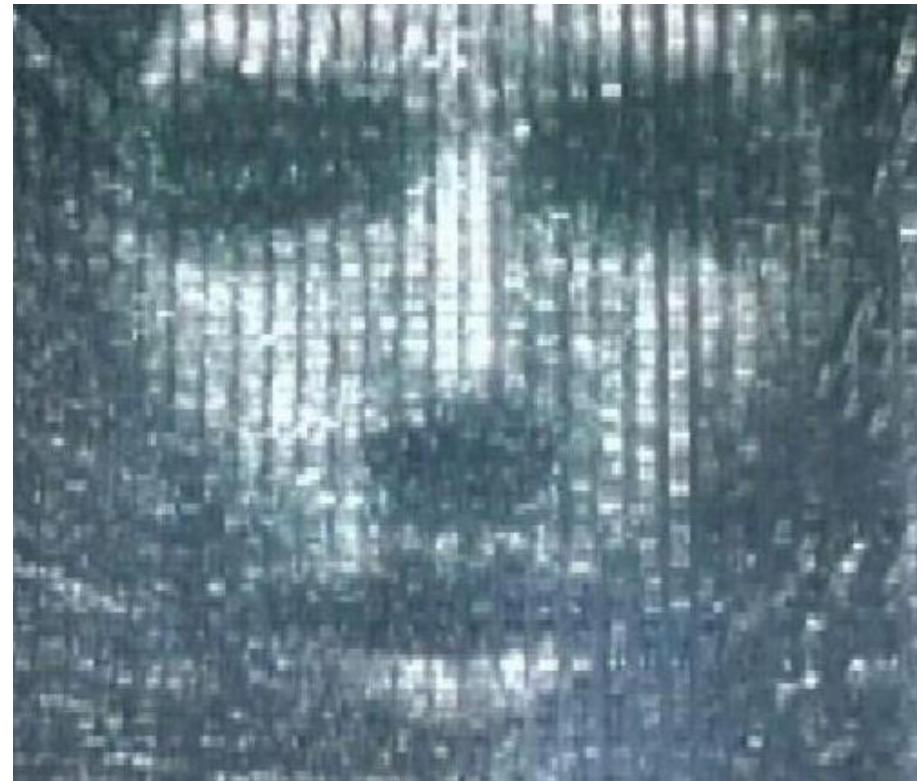
Natural Language Generation: Task

Antoine Bosselut



What is natural language generation?

- Natural language generation (NLG) is a sub-field of natural language processing
- Focused on building systems that automatically produce **coherent** and **useful** written or spoken text for human consumption
- NLG systems are already changing the world we live in...



Machine Translation



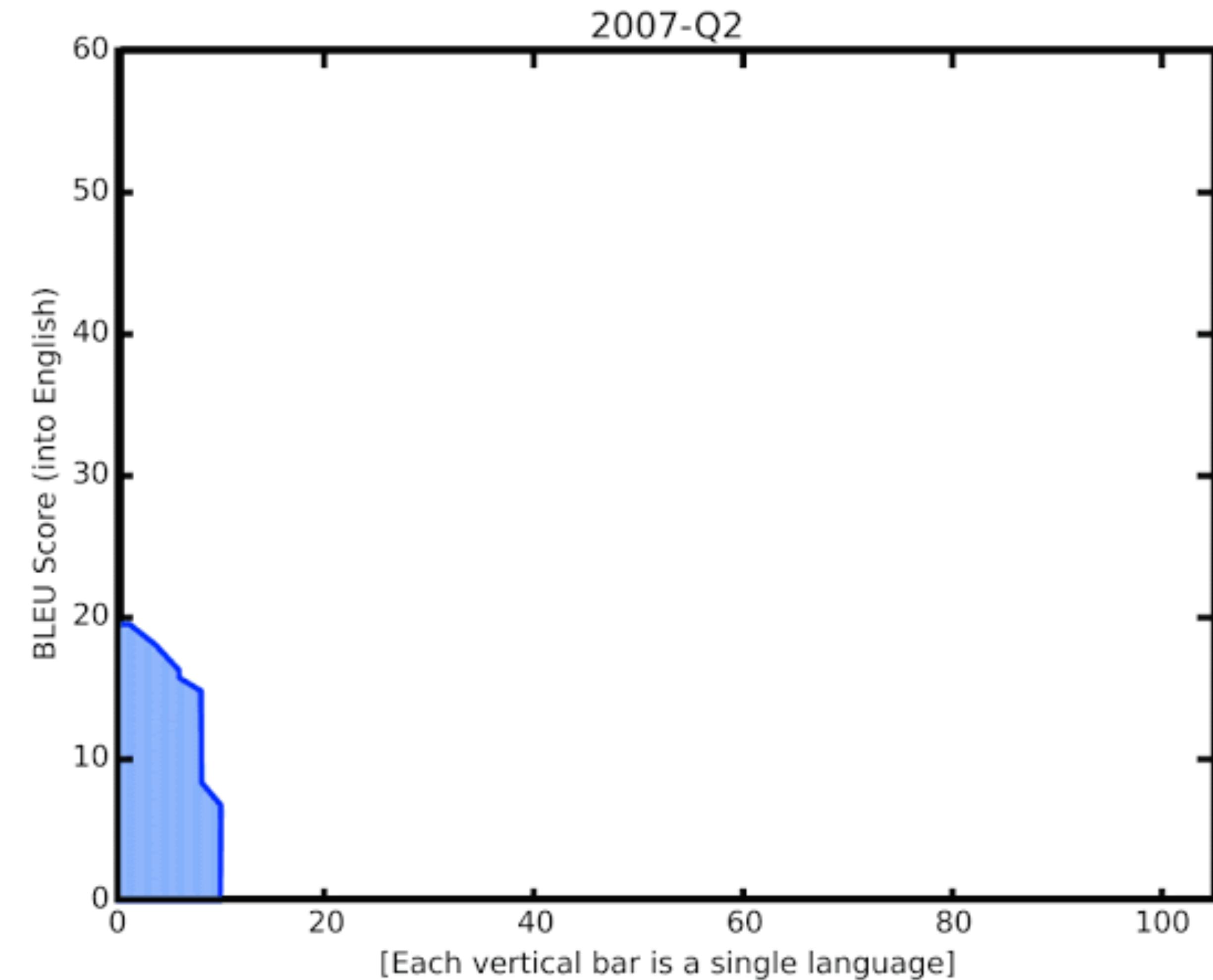
French ▾ ↗ English ▾

J'ai mangé avec
mon avocat
aujourd'hui

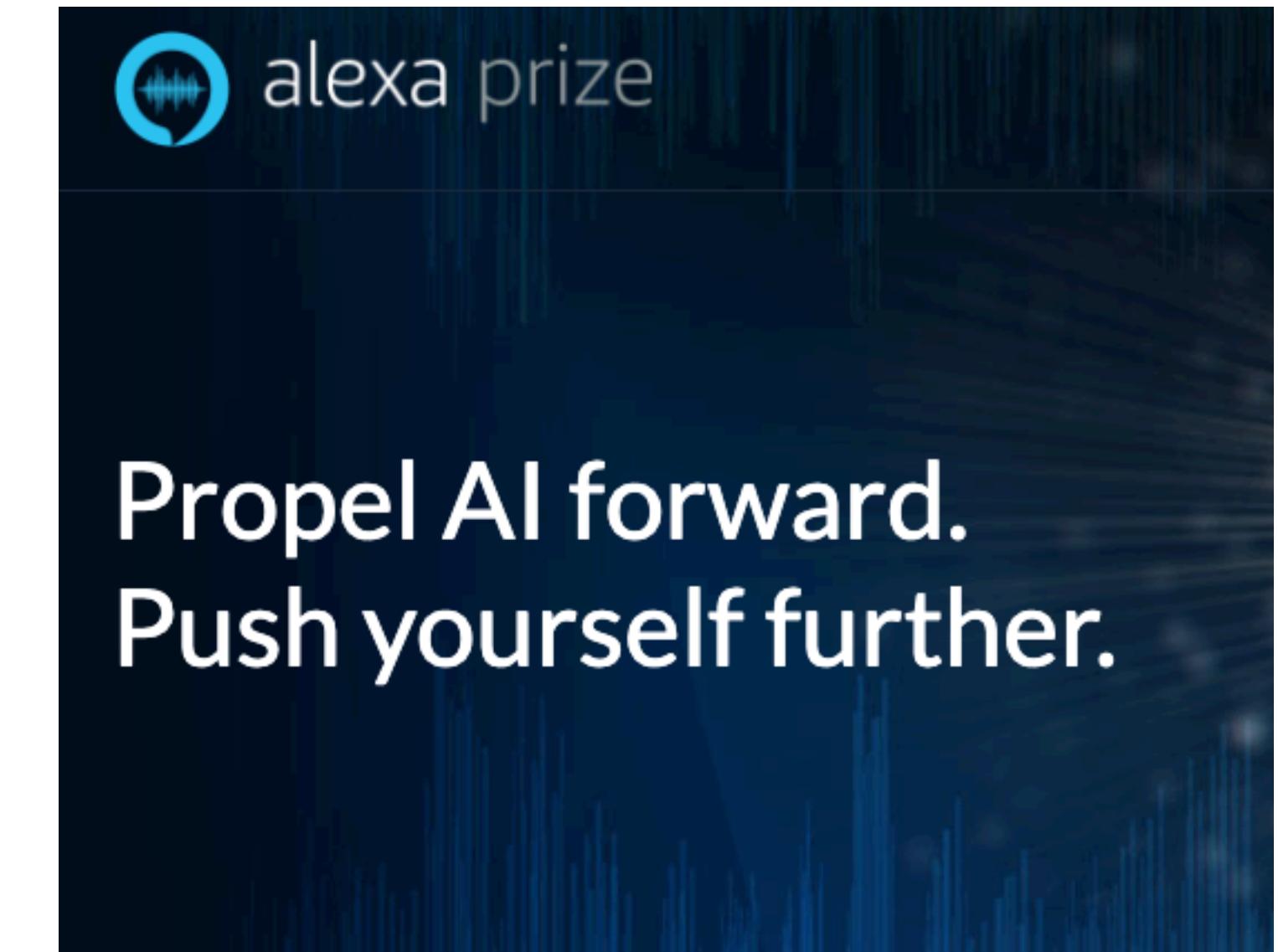
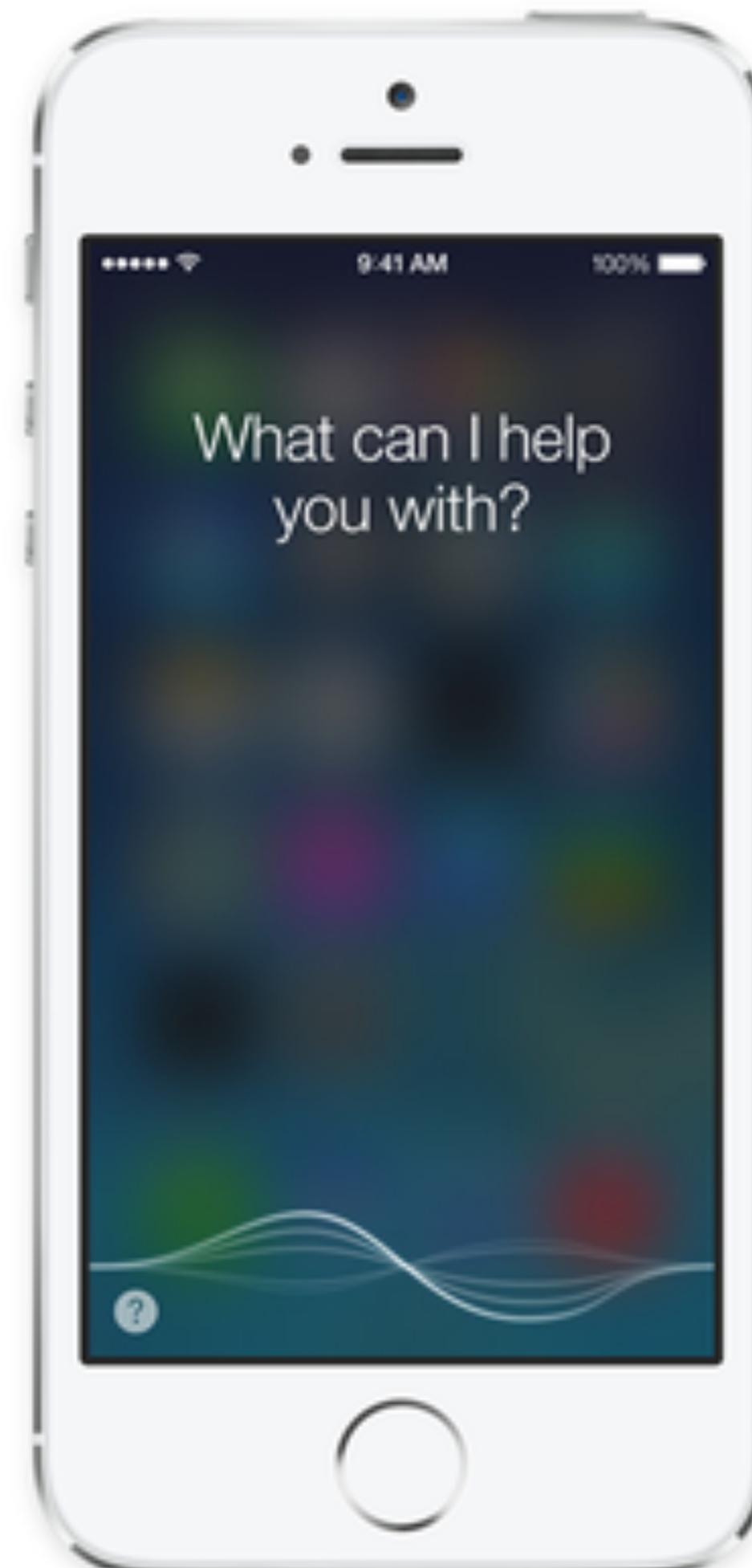
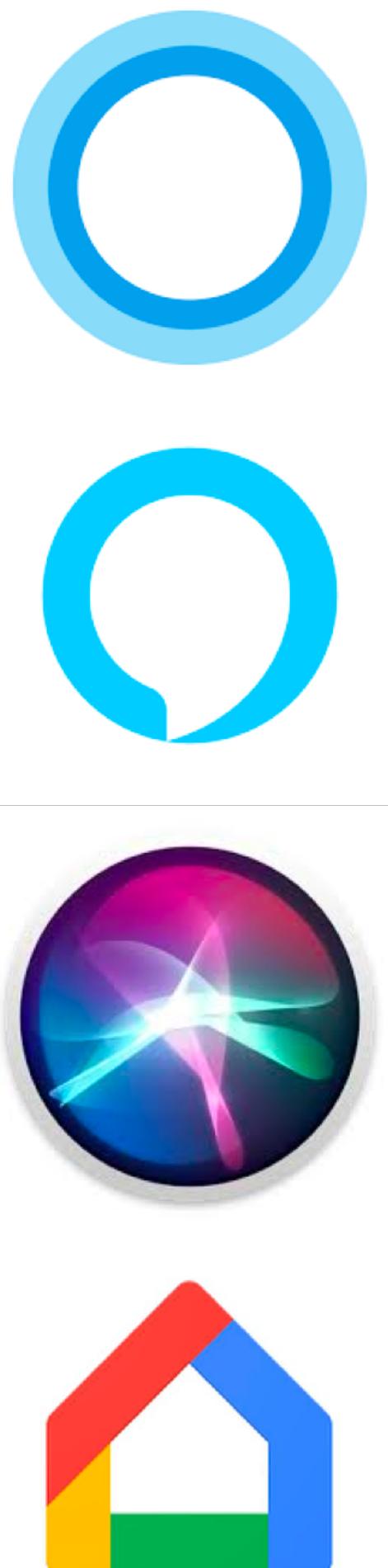
I ate with my lawyer
today

×

Speaker icon Microphone icon Speaker icon Copy icon

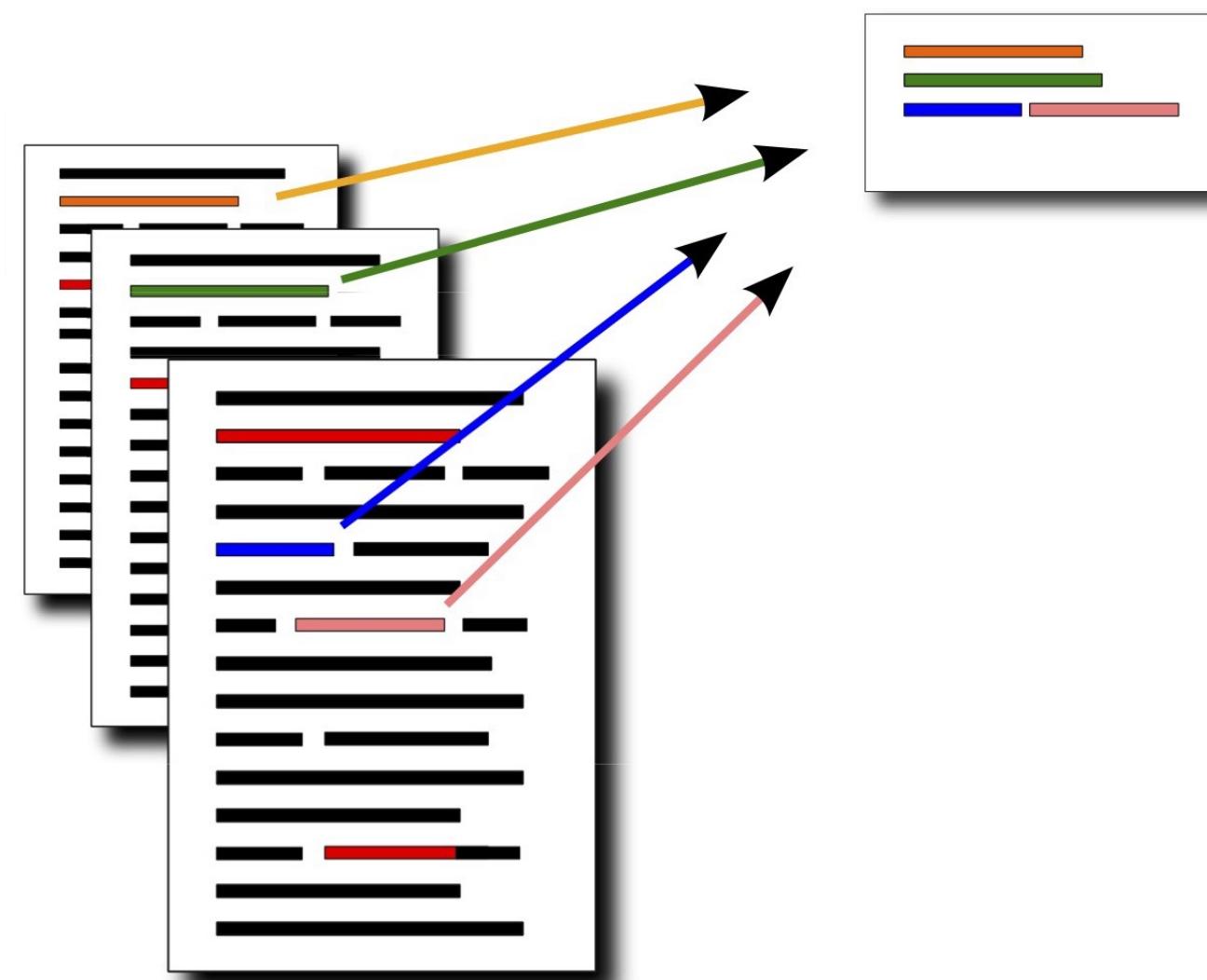
A screenshot of a machine translation interface. It shows a French sentence "J'ai mangé avec mon avocat aujourd'hui" on the left and its English translation "I ate with my lawyer today" on the right. The interface includes language selection dropdowns, a central bidirectional arrow, and standard copy/paste controls at the bottom.

Dialogue Systems



Summarization

Document Summarization



<http://mogren.one/lic/>

E-mail Summarization

re-thinking com.cy—1 min read, 122 words

TL;DR: Anyone should be able to buy a .cy domain regardless of location, in a quick and efficient way

1 min read, 122 words

Argyrou Argyris <argyrou.a@gmail.com> to me Sep 8, 2019, 11:53 AM

Hey,

Cyprus country code TLD registrar [nic.cy](#) operated by the University of Cyprus is the ONLY way to register a [com.cy](#) domain in Cyprus. We are talking about a bureaucratic process.

I still don't get it why we can't freely register .cy names. Right now you can't buy .cy domains, only [com.cy](#), and a list of other [whatever-useless.cy](#) domain extensions.

Releasing .cy will help the sales and promotion of our national country code top level domain. It will be a new domain introduced on the web and therefore many available names will be free to register. **Anyone should be able to buy a .cy domain regardless of location, in a quick and efficient way.**

[nic.cy](#) should provide this exclusive domain to registrars and their customers worldwide.

<https://chrome.google.com/webstore/detail/gmail-summarization/>

Meeting Summarization

C: Looking at what we've got, we want an LCD display with a spinning wheel.
B: You have to have some push-buttons, don't you?
C: Just spinning and not scrolling, I would say.
B: I think the spinning wheel is definitely very now.
A: but since LCDs seems to be uh a definite yes,
C: We're having push-buttons on the outside
C: and then on the inside an LCD with spinning wheel,

Decision Abstract (Summary):
The remote will have push buttons outside, and an LCD and spinning wheel inside.

A: and um I'm not sure about the buttons being in the shape of fruit though.
D: Maybe make it like fruity colours or something.
C: The power button could be like a big apple or something.
D: Um like I'm just thinking bright colours.

Problem Abstract (Summary):
How to incorporate a fruit and vegetable theme into the remote.

(Wang and Cardie, ACL 2013)

Data-to-Text Generation

Table Title: Robert Craig (American football)
Section Title: National Football League statistics
Table Description: None

YEAR	TEAM	RUSHING					RECEIVING				
		ATT	YDS	AVG	LNG	TD	NO.	YDS	AVG	LNG	TD
1983	SF	176	725	4.1	71	8	48	427	8.9	23	4
1984	SF	155	649	4.2	28	4	71	675	9.5	64	3
1985	SF	214	1050	4.9	62	9	92	1016	11	73	6
1986	SF	204	830	4.1	25	7	81	624	7.7	48	0
1987	SF	215	815	3.8	25	3	66	492	7.5	35	1
1988	SF	310	1502	4.8	46	9	76	534	7.0	22	1
1989	SF	271	1054	3.9	27	6	49	473	9.7	44	1
1990	SF	141	439	3.1	26	1	25	201	8.0	31	0
1991	RAI	162	590	3.6	15	1	17	136	8.0	20	0
1992	MIN	105	416	4.0	21	4	22	164	7.5	22	0
1993	MIN	38	119	3.1	11	1	19	169	8.9	31	1
Totals	-	1991	8189	4.1	71	56	566	4911	8.7	73	17

Target Text: Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

(Parikh et al., EMNLP 2020)

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS ...
Heat	11	12	103	49	47	27
Hawks	7	15	95	43	33	20
PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta

The Atlanta Hawks defeated the Miami Heat , 103 - 95 , at Philips Arena on Wednesday . Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here . Defense was key for the Hawks , as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers . Atlanta also dominated in the paint , winning the rebounding battle , 47 - 34 , and outscoring them in the paint 58 - 26.The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets . This was a near wire - to - wire win for the Hawks , as Miami held just one lead in the first five minutes . Miami (7 - 15) are as beat - up as anyone right now and it 's taking a toll on the heavily used starters . Hassan Whiteside really struggled in this game , as he amassed eight points , 12 rebounds and one blocks on 4 - of - 12 shooting ...

MR:

name[The Eagle],
eatType[coffee shop],
food[French],
priceRange[moderate],
customerRating[3/5],
area[riverside],
kidsFriendly[yes],
near[Burger King]

NL:

"The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King."

(Wiseman and Rush., EMNLP 2017)

(Dusek et. al., INLG 2019)

Visual Description Generation



bowls are food in triangular shape are sitting on table
table filled with many plates of various breakfast foods
table topped with lots of different types of donuts



hotdog stand on busy street
man in white t shirt is holding umbrella and ice cream cart
man in white shirt is pushing his cart down street



man in graduation robes riding bicycle
cyclist giving thumbs up poses with his bicycle by right of way sign at park
man riding motorcycle on street



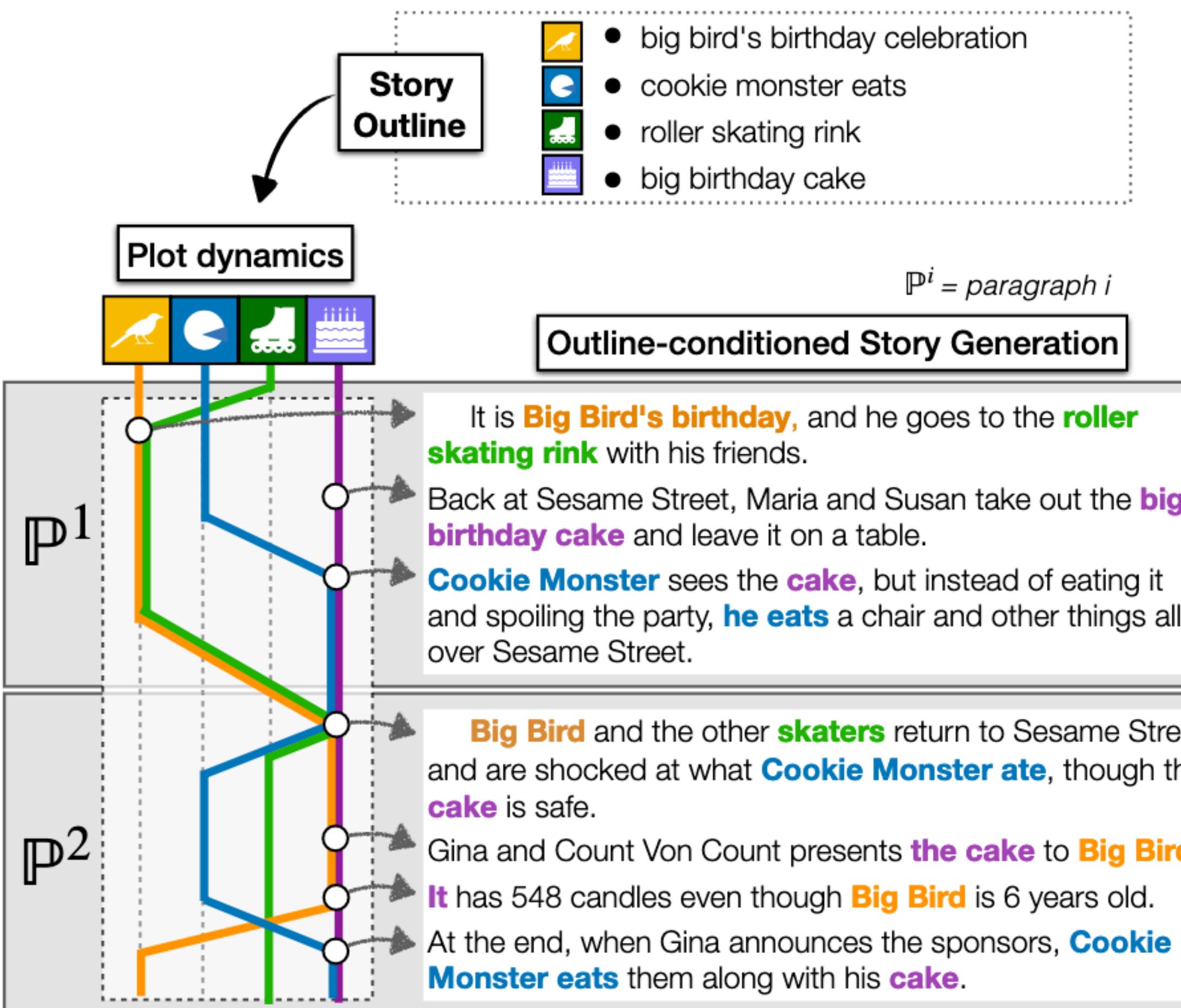
one man and two women sitting in living room
man and woman are playing wii game while woman sits on couch with wine glass in her hand
group of people sitting on couch with their laptops



Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

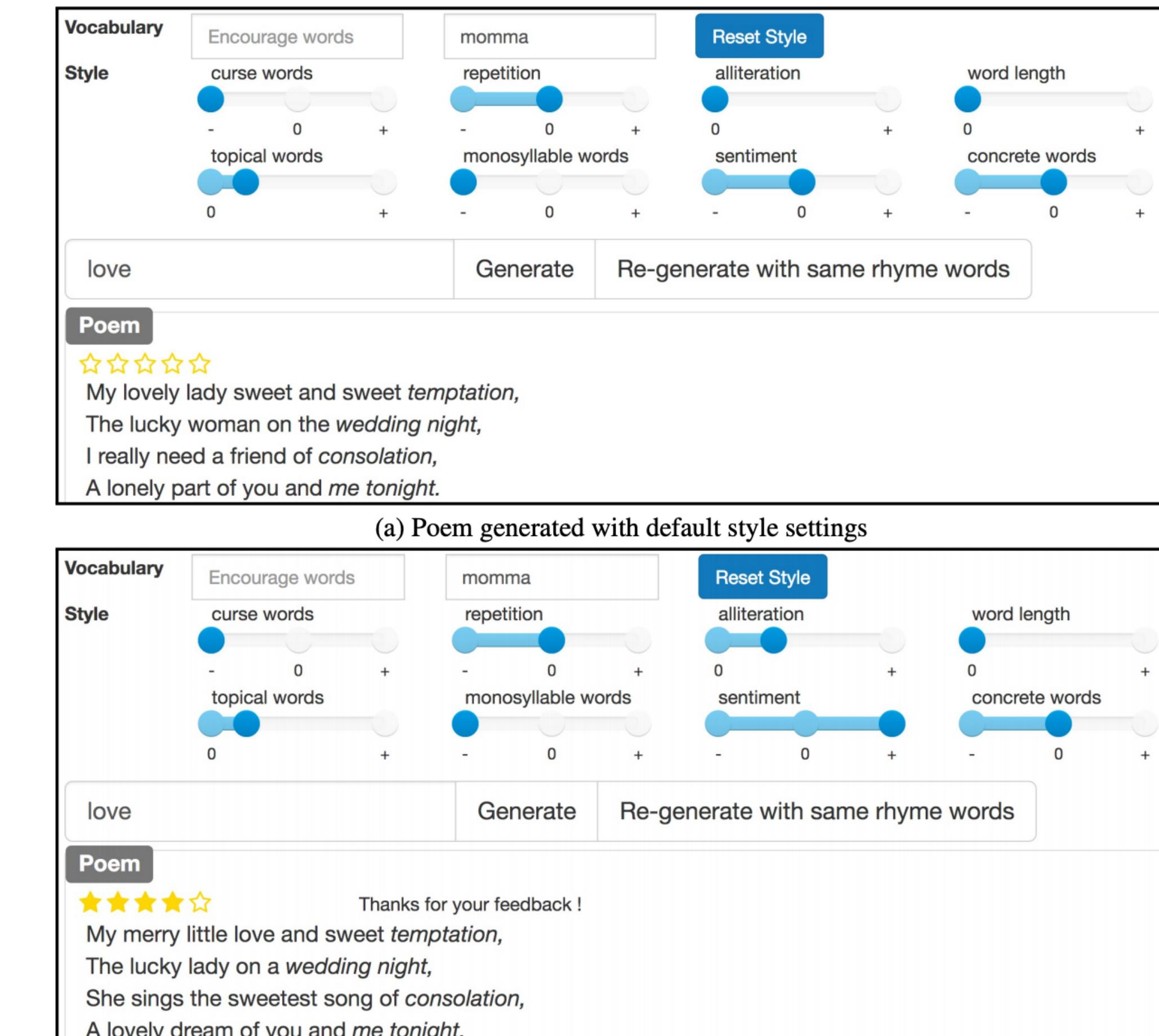
Creative Generation

Stories & Narratives



(Rashkin et al., EMNLP 2020)

Poetry



(b) Poem generated with user adjusted style settings

(Ghazvininejad et al., ACL 2017)

All-in-one: ChatGPT

AN

Can you give me a two-day itinerary for a weekend trip to Paris ? Please don't include more than three activities per day.



Day 1:

1. Visit the Eiffel Tower - Start your day with the iconic symbol of Paris, the Eiffel Tower. Take a stroll around the Champ de Mars and enjoy the views from the tower.
2. Explore the Louvre Museum - After lunch, head to the Louvre Museum and admire the world-famous artwork such as the Mona Lisa and the Venus de Milo.
3. Stroll along the Seine River - In the evening, take a romantic stroll along the Seine River and enjoy the Parisian atmosphere.

Day 2:

1. Visit Notre-Dame Cathedral - Start your day by visiting one of the most famous Gothic cathedrals in the world, Notre-Dame Cathedral.
2. Explore the Marais neighborhood - After lunch, explore the Marais neighborhood, known for its historical architecture, trendy boutiques, and restaurants.
3. Enjoy a Cabaret Show - In the evening, enjoy a cabaret show at the famous Parisian venue, the Moulin Rouge, and end your trip with a memorable experience.

What is natural language generation?

Any task involving text production for human consumption requires natural language generation

What is natural language generation?

Any task involving text production for human consumption requires natural language generation

Deep Learning is powering next-gen NLG systems!

Today's Outline

- **Introduction**
- **Section 1:** Formalizing NLG: a simple model for training and decoding
- **Next week:** Deeper dive into training, decoding and evaluation
- **Exercise Session:** Robustness & Prompting

Basics of natural language generation

- Most text generation are autoregressive models — they predict **next tokens** based on the values of **past tokens**
- In autoregressive text generation models, at each time step t , our model takes in a sequence of tokens of text as input $\{y\}_{<t}$ and outputs a new token, \hat{y}_t

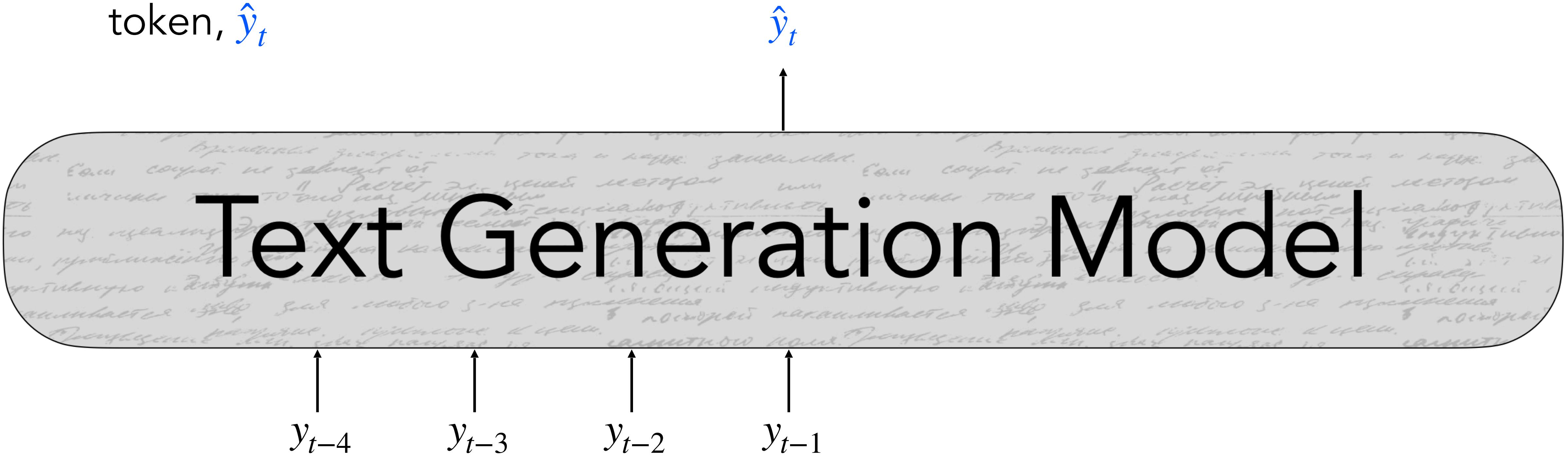
Basics of natural language generation

- In autoregressive text generation models, at each time step t , our model takes in a sequence of tokens of text as input $\{y\}_{<t}$ and outputs a new token, \hat{y}_t

Text Generation Model

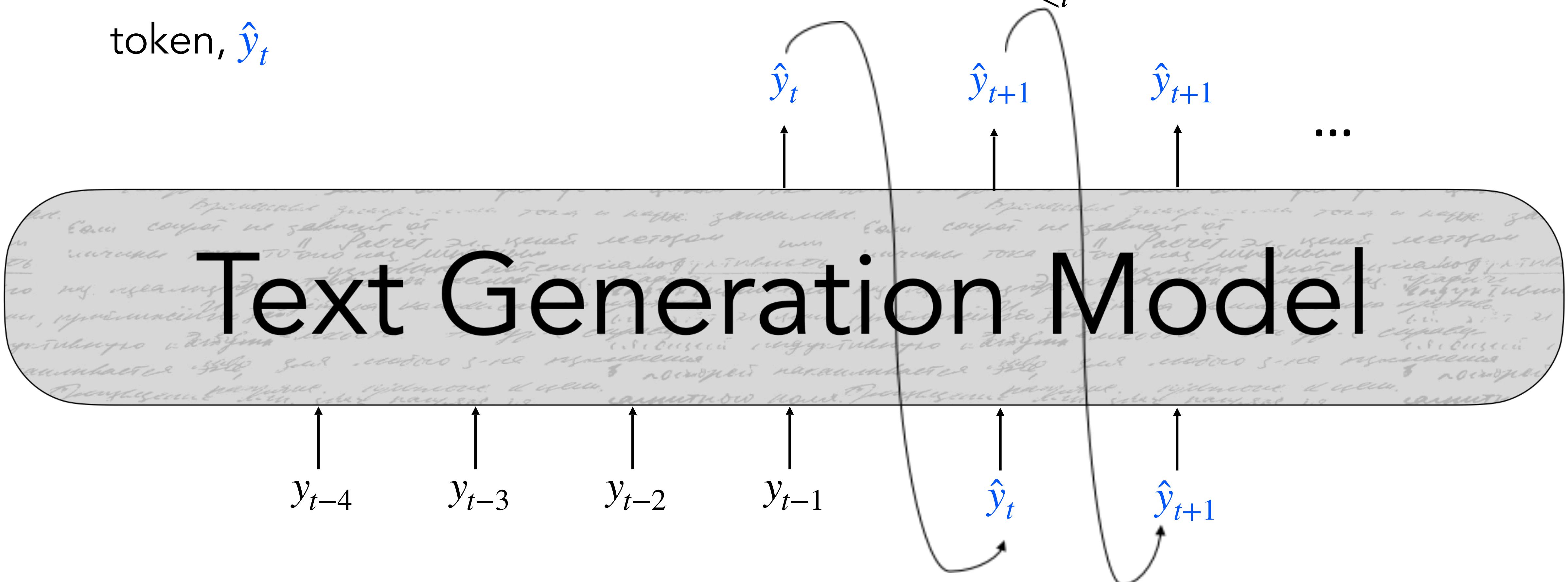
Basics of natural language generation

- In autoregressive text generation models, at each time step t , our model takes in a sequence of tokens of text as input $\{y\}_{<t}$ and outputs a new token, \hat{y}_t



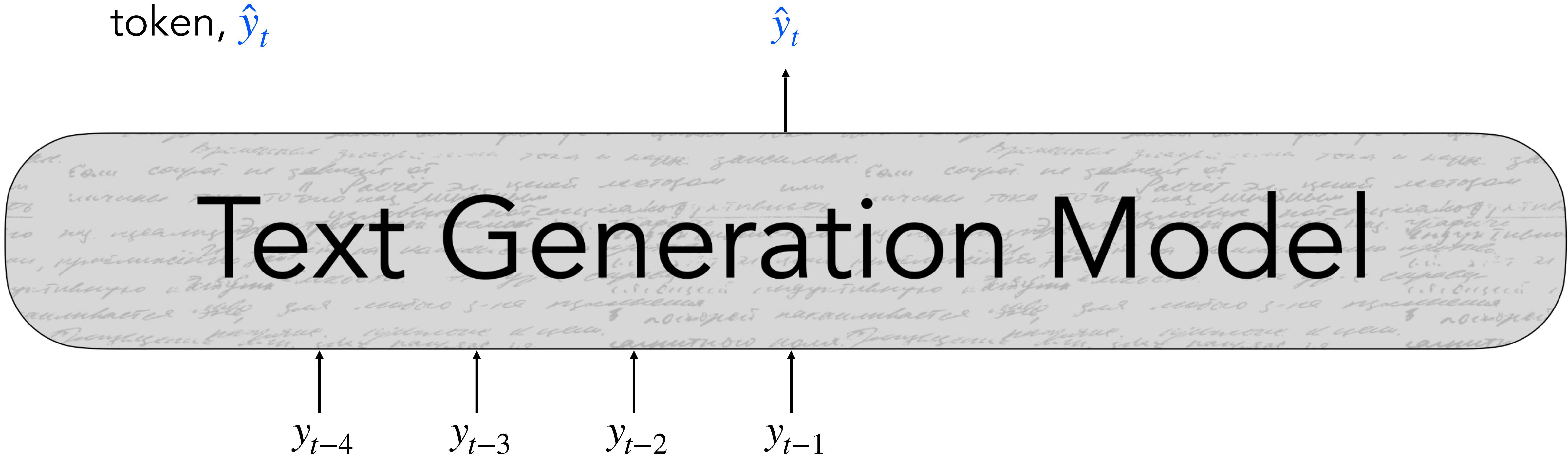
Basics of natural language generation

- In autoregressive text generation models, at each time step t , our model takes in a sequence of tokens of text as input $\{y\}_{<t}$ and outputs a new token, \hat{y}_t



A look at a single step

- In autoregressive text generation models, at each time step t , our model takes in a sequence of tokens of text as input $\{y\}_{<t}$ and outputs a new token, \hat{y}_t



A look at a single step

- At each time step t , our model computes a vector of scores for each token in our vocabulary, $\mathbf{S} \in \mathbb{R}^V$:

$$\mathbf{S} = f\left(\{y_{<t}\}, \theta\right)$$

$f(\cdot)$ is your model

- Then, we compute a probability distribution \mathbf{P} over $w \in V$ using these scores:

$$P(y_t = w \mid \{y_{<t}\}) = \frac{\exp(S_w)}{\sum_{w' \in V} \exp(S_{w'})}$$

A look at a single step

- At each time step t , our model computes a vector of scores for each token in our vocabulary, $\mathbf{S} \in \mathbb{R}^V$:

$$\mathbf{S} = f\left(\{y_{<t}\}, \theta\right)$$

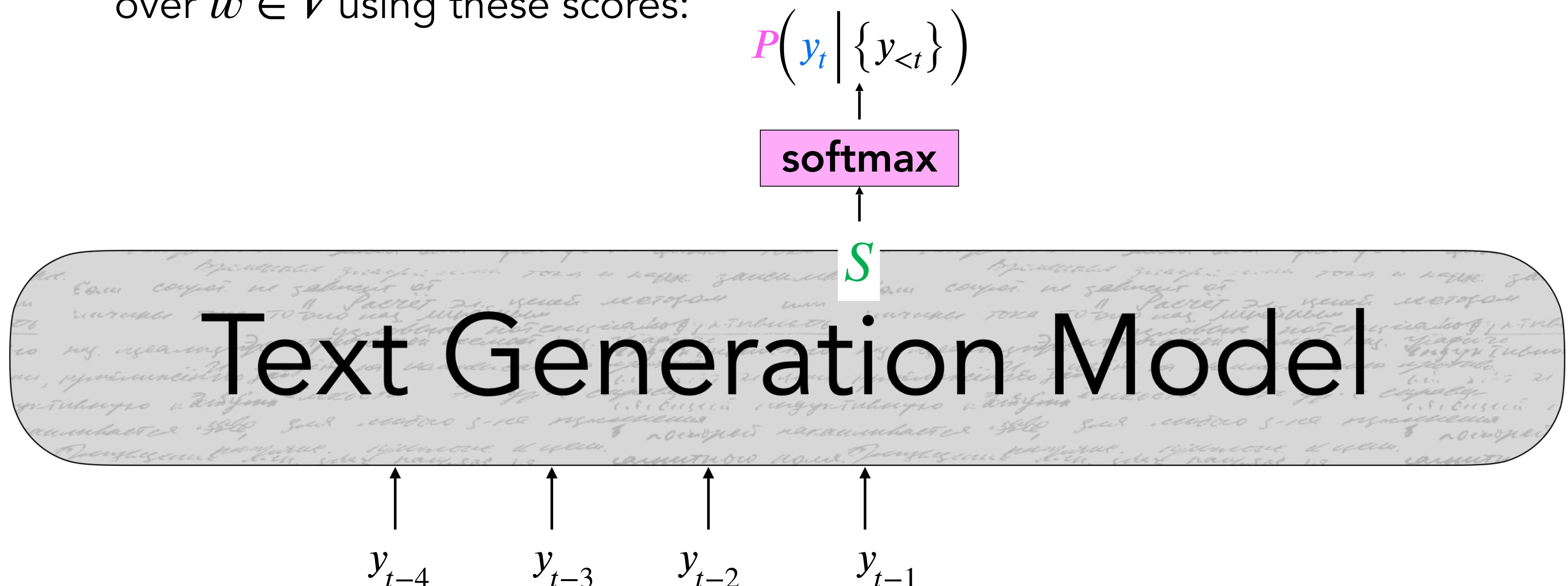
$f(\cdot)$ is your model

- Then, we compute a probability distribution \mathbf{P} over $w \in V$ using these scores:

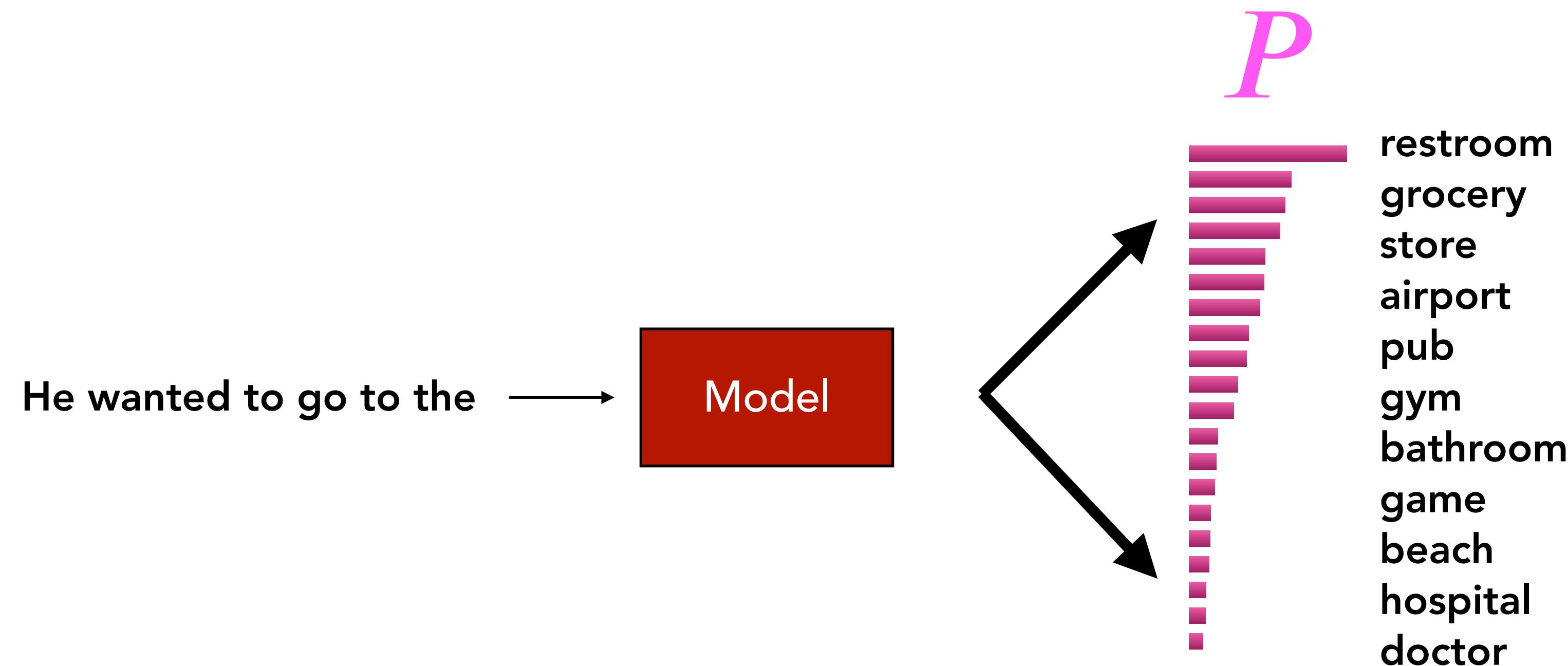
$$P(y_t | \{y_{<t}\}) = \frac{\exp(S_w)}{\sum_{w' \in V} \exp(S_{w'})}$$

A look at a single step

- At each time step t , our model computes a vector of scores for each token in our vocabulary, $\mathbf{S} \in \mathbb{R}^V$. Then, we compute a probability distribution \mathbf{P} over $w \in V$ using these scores:



A look at a single step



- At inference time, our decoding algorithm defines a function to select a token from this distribution P :

$$\hat{y}_t = g(P(y_t | \{y_{<t}\}))$$

g(.) is your decoding algorithm

Basics: What are we trying to do?

- We train the model to minimize the negative loglikelihood of predicting the next token in the sequence:

$$\mathcal{L}_t = -\log P(y_t^* | \{y_{<t}^*\})$$

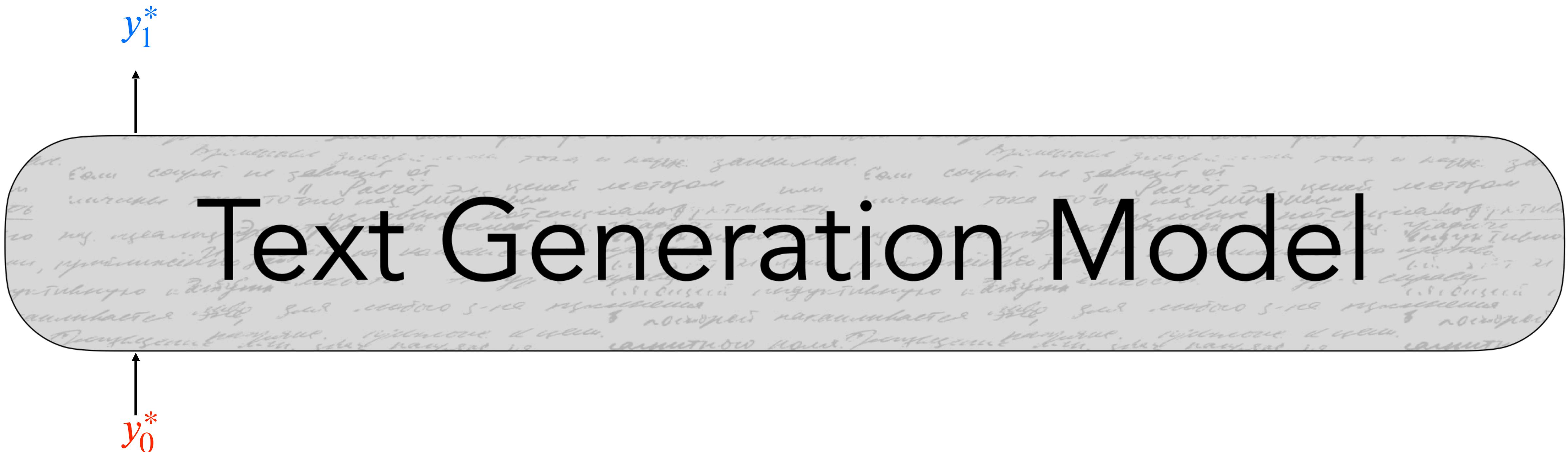
Sum \mathcal{L}_t for the entire sequence

- This is a **multi-class classification task** where each $w \in V$ is a unique class.
- The label at each step is the actual word y_t^* in the training sequence
- This token is often called the “**gold**” or “**ground truth**” token
- This algorithm is often called “teacher forcing”

Maximum Likelihood Training (i.e., *teacher forcing*)

- Trained to generate the next word y_t^* given a set of preceding words $\{y^*\}_{<t}$

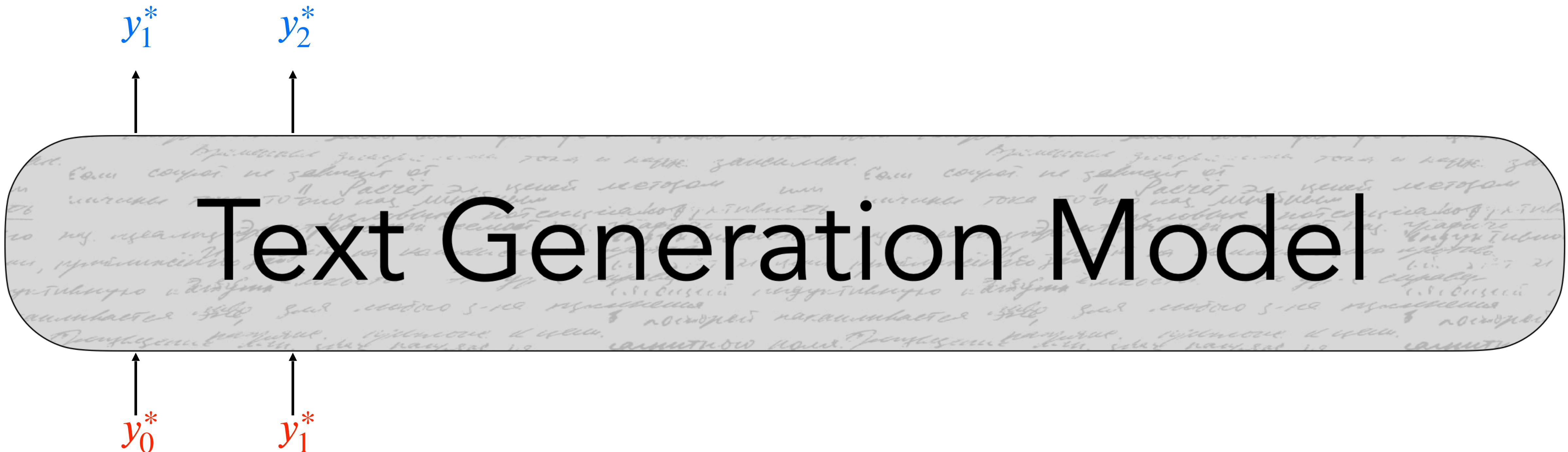
$$\mathcal{L} = -\log P(y_1^* | y_0^*)$$



Maximum Likelihood Training (i.e., *teacher forcing*)

- Trained to generate the next word y_t^* given a set of preceding words $\{y^*\}_{<t}$

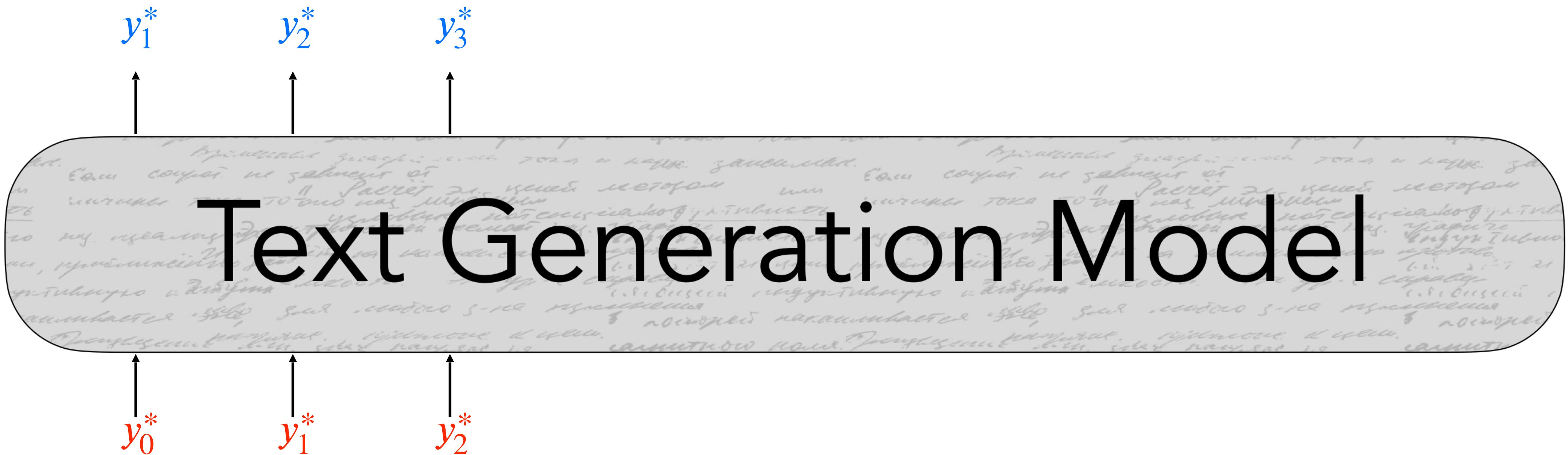
$$\mathcal{L} = -(\log P(y_1^* | y_0^*) + \log P(y_2^* | y_0^*, y_1^*))$$



Maximum Likelihood Training (i.e., *teacher forcing*)

- Trained to generate the next word y_t^* given a set of preceding words $\{y^*\}_{<t}$

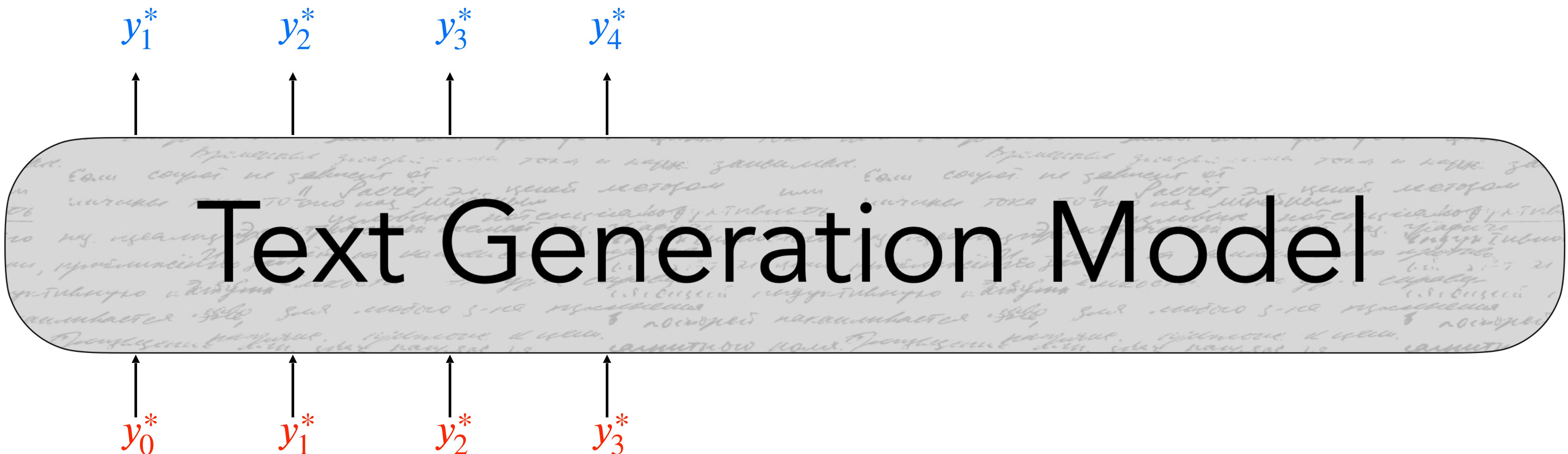
$$\mathcal{L} = - (\log P(y_1^* | y_0^*) + \log P(y_2^* | y_0^*, y_1^*) + \log P(y_3^* | y_0^*, y_1^*, y_2^*))$$



Maximum Likelihood Training (i.e., *teacher forcing*)

- Trained to generate the next word y_t^* given a set of preceding words $\{y^*\}_{<t}$

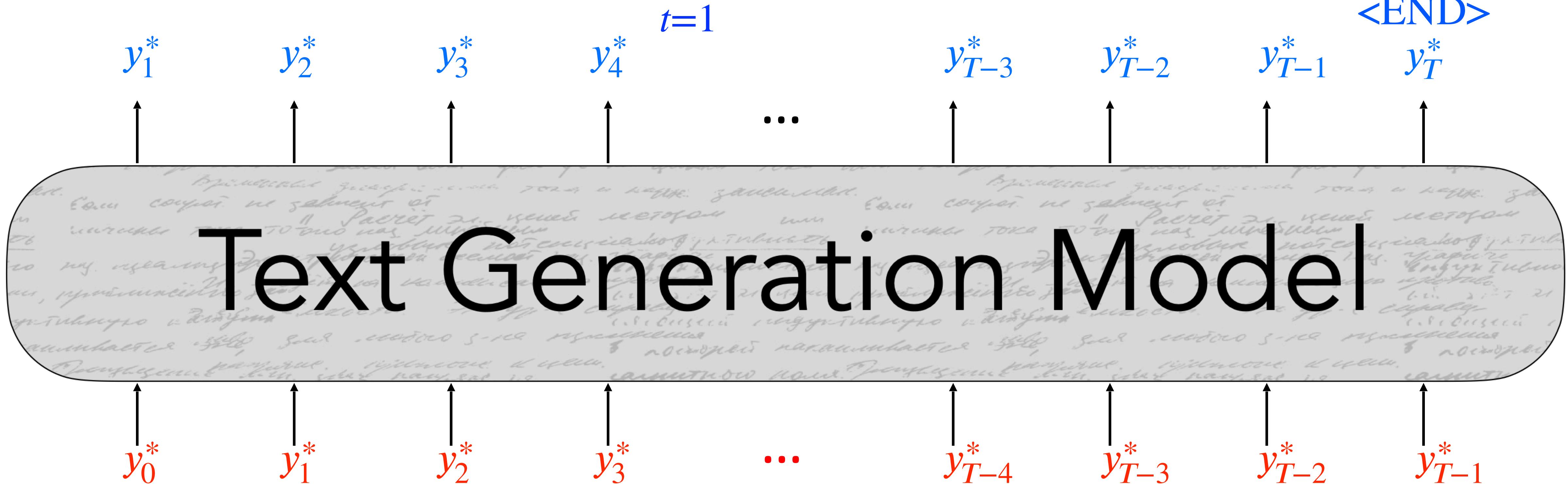
$$\mathcal{L} = - \sum_{t=1}^4 \log P(y_t^* | \{y^*\}_{<t})$$



Maximum Likelihood Training (i.e., *teacher forcing*)

- Trained to generate the next word y_t^* given a set of preceding words $\{y_{<t}^*\}$

$$\mathcal{L} = - \sum_{t=1}^T \log P(y_t^* | \{y_{<t}^*\})$$



Text Generation: Takeaways

- Text generation is the foundation of many useful NLP applications (e.g., translation, summarisation, dialogue systems)
- In autoregressive NLG, we generate one token a time, using the context and previous generated tokens as inputs for generating the next token.
- Our model generates a set of scores for every token in the vocabulary, which we can convert to a probability distribution using the softmax function
- To get a calibrated distribution, we train our model using maximum likelihood estimation to predict the next token on a dataset of sequences

Decoding: what is it all about?

- At each time step t , our model computes a vector of scores for each token in our vocabulary, $\mathbf{S} \in \mathbb{R}^V$:

$$\mathbf{S} = f(\{y_{<t}\})$$

$f(\cdot)$ is your model

- Then, we compute a probability distribution \mathbf{P} over these scores (usually with a softmax function):

$$P(y_t = w | \{y_{<t}\}) = \frac{\exp(S_w)}{\sum_{w' \in V} \exp(S_{w'})}$$

- Our decoding algorithm defines a function to select a token from this distribution:

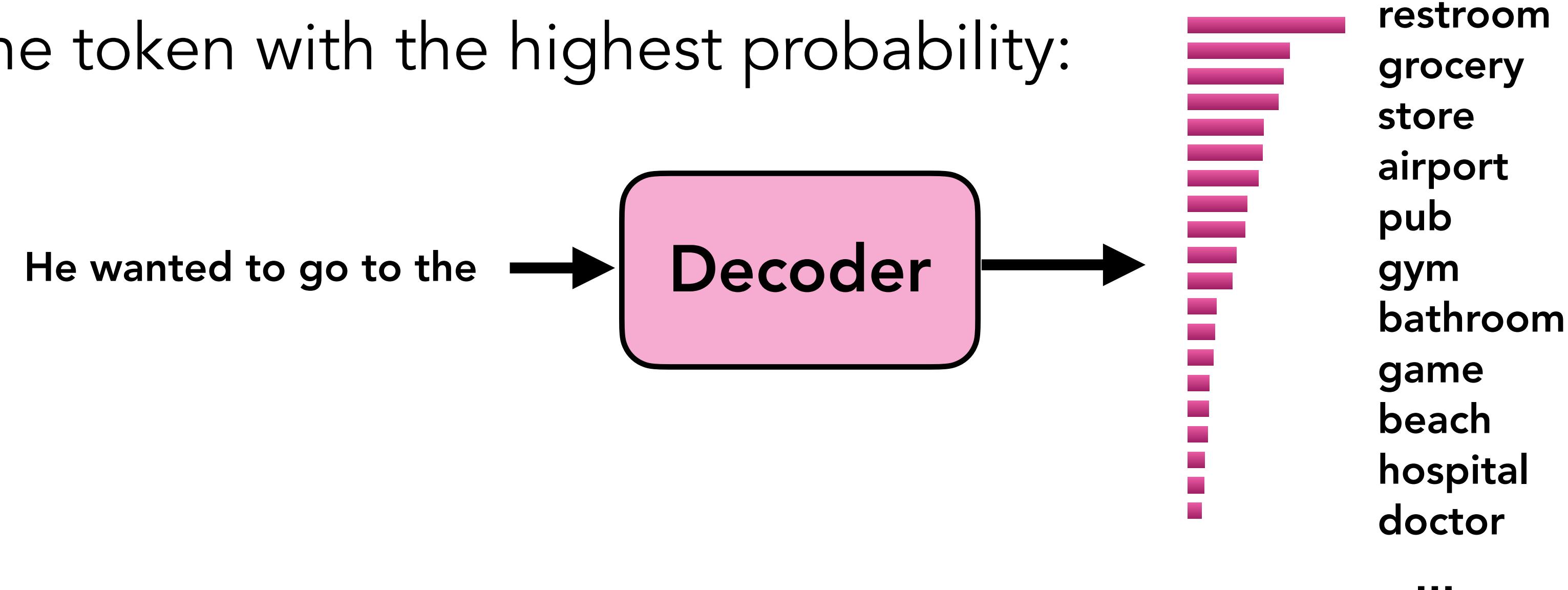
$$\hat{y}_t = g(P(y_t | \{y_{<t}\}))$$

$g(\cdot)$ is your decoding algorithm

Greedy methods: Argmax Decoding

$$\hat{y}_t = \underset{w \in V}{\operatorname{argmax}} P(y_t = w | \{y\}_{<t})$$

- g = select the token with the highest probability:



Greedy methods: Argmax Decoding

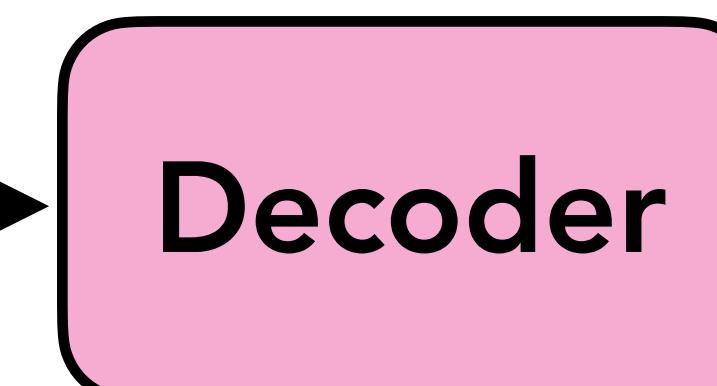
$$\hat{y}_t = \text{argmax } P(y_t = w | \{y\}_{\leq t})$$

Select highest scoring token

What's a potential problem with argmax decoding?

- g = selected

He wanted to go to the



store
airport
pub
gym
bathroom
game
beach
hospital
doctor
...

Issues with argmax decoding

- In argmax decoding, we cannot revise prior decisions
 - *les pauvres sont démunis (the poor don't have any money)*
 - → *the _____*
 - → *the poor _____*
 - → *the poor **are** _____*
- Potential leads to sequences that are
 - **Ungrammatical**
 - **Unnatural**
 - **Nonsensical**
 - **Incorrect**

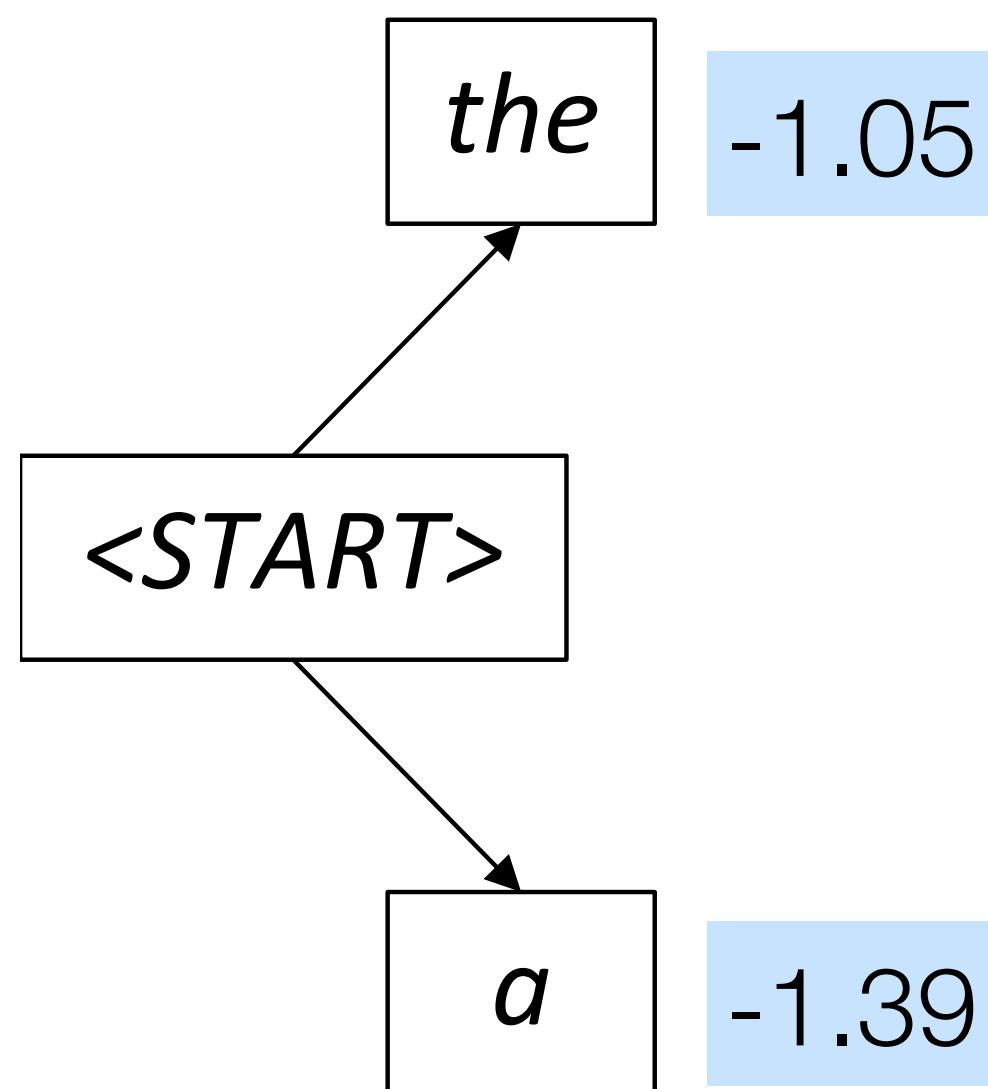
Beam Search

- *les pauvres sont démunis (the poor don't have any money)*
- → *the _____*
- → *the poor _____*
- → *the poor **are** _____*
- **Beam Search:** Explore several different hypotheses instead of just one
 - Track of the b highest scoring sequences at each decoder step instead of just one
 - Score at each step:
$$\sum_{t=1}^j \log \textcolor{blue}{P}(\hat{y}_t | \hat{y}_1, \dots, \hat{y}_{t-1}, X)$$
 - b is called the **beam size**

Beam Search

Beam size = 2

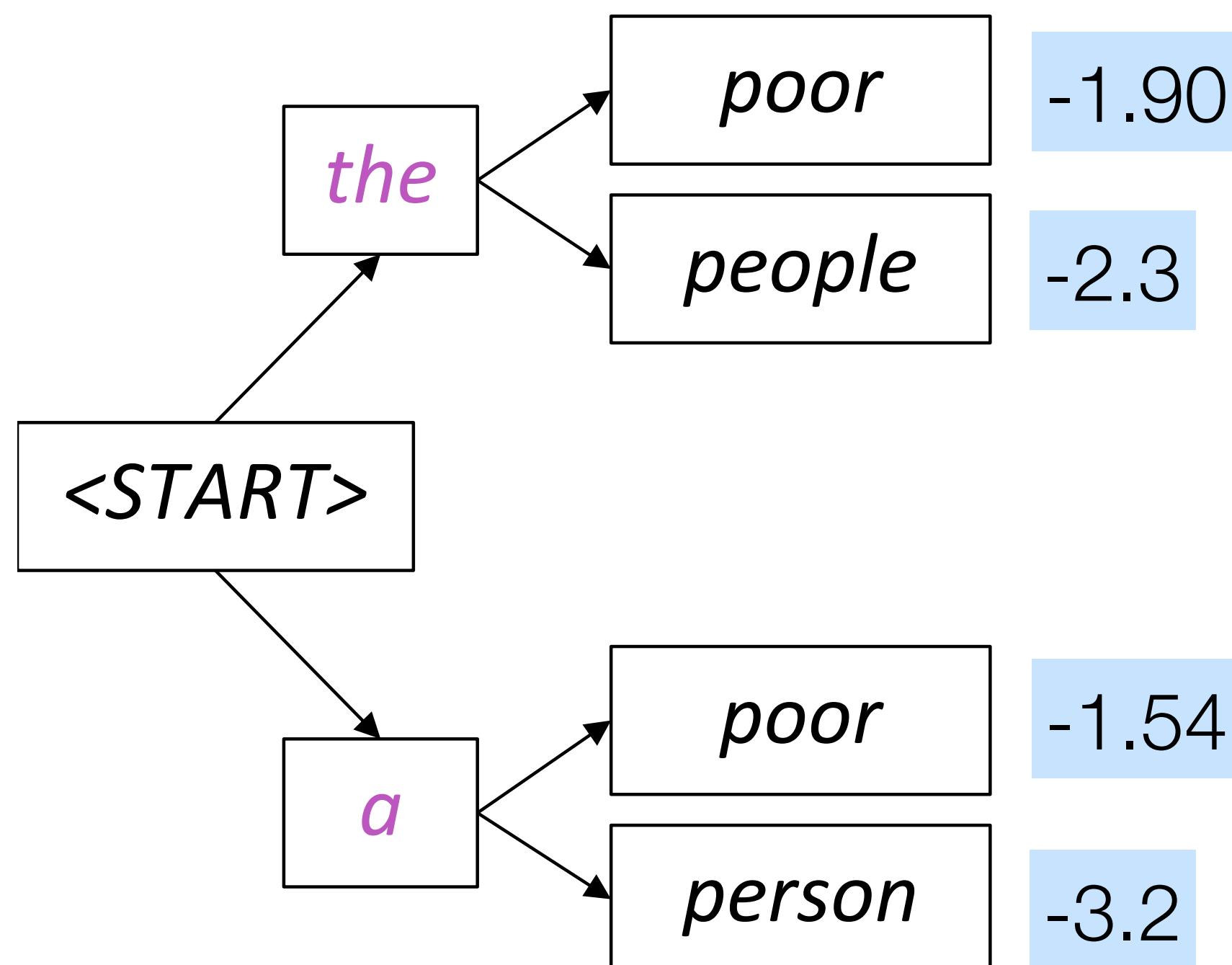
$$\log \textcolor{magenta}{P}(\hat{y}_1 | y_0)$$



Beam Search

Beam size = 2

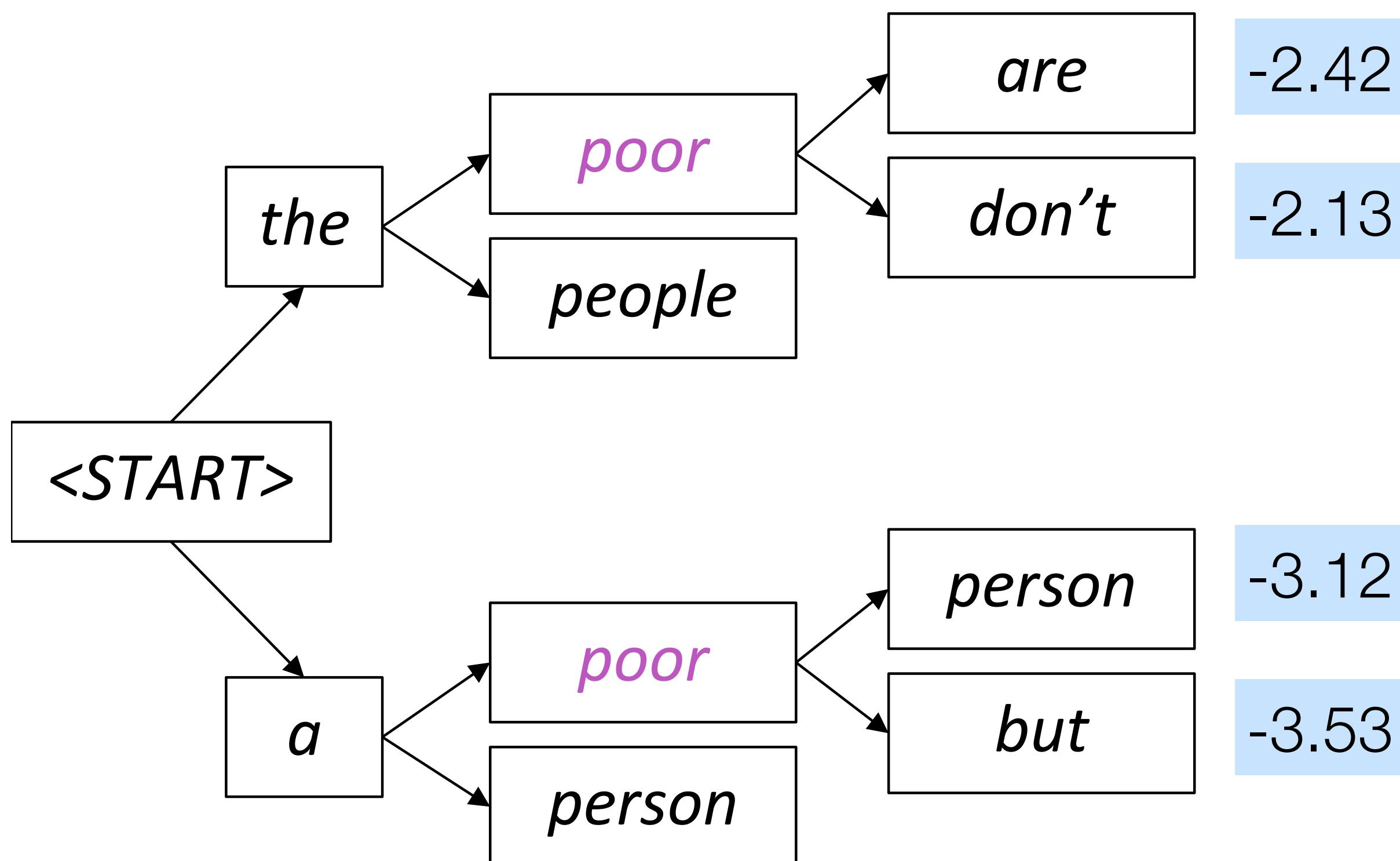
$$\sum_{t=1}^2 \log P(\hat{y}_t | \hat{y}_0, \dots, \hat{y}_{t-1})$$



Beam Search

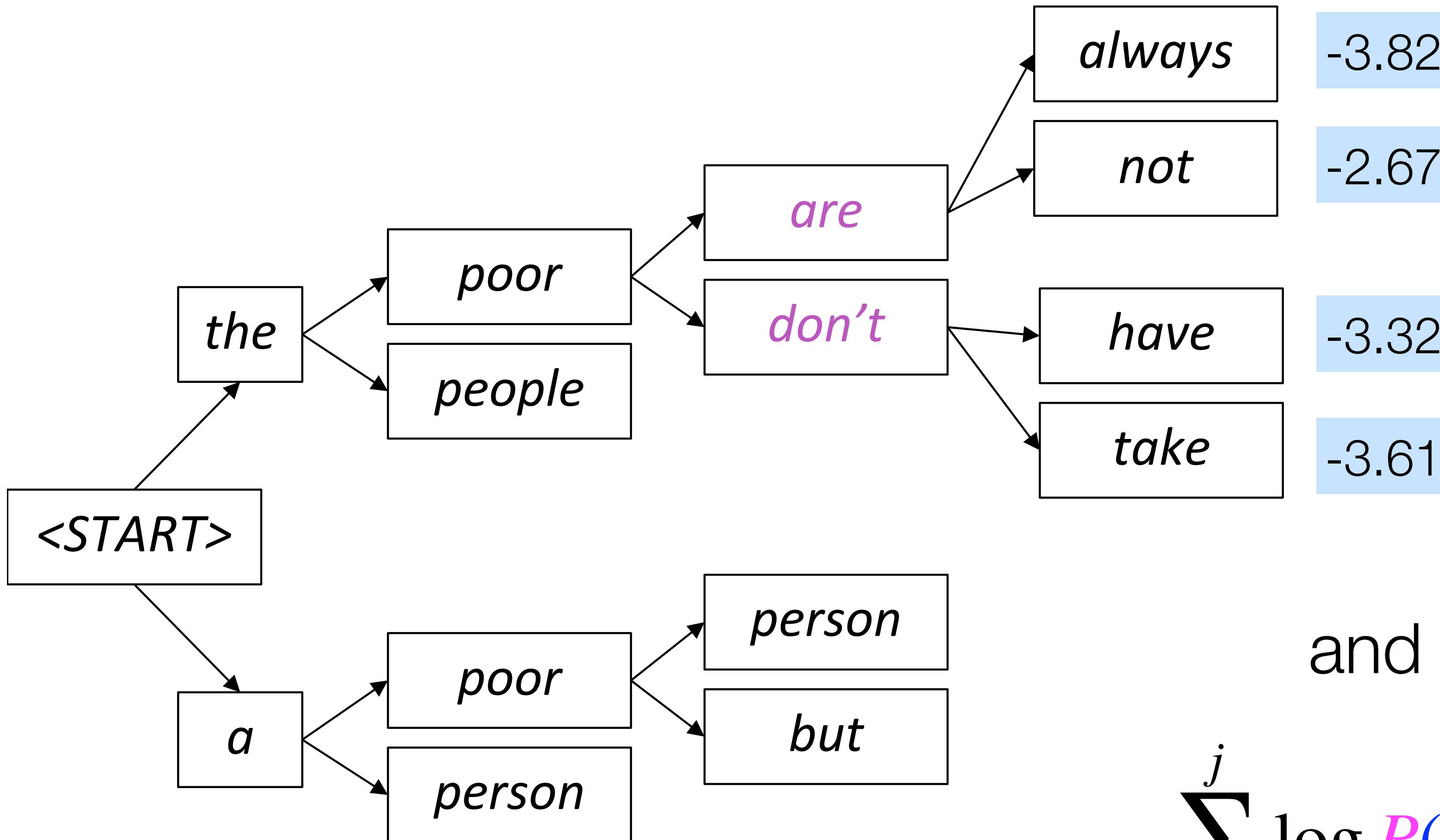
Beam size = 2

$$\sum_{t=1}^3 \log P(\hat{y}_t | y_0, \hat{y}_1, \dots, \hat{y}_{t-1})$$



Beam Search

Beam size = 2

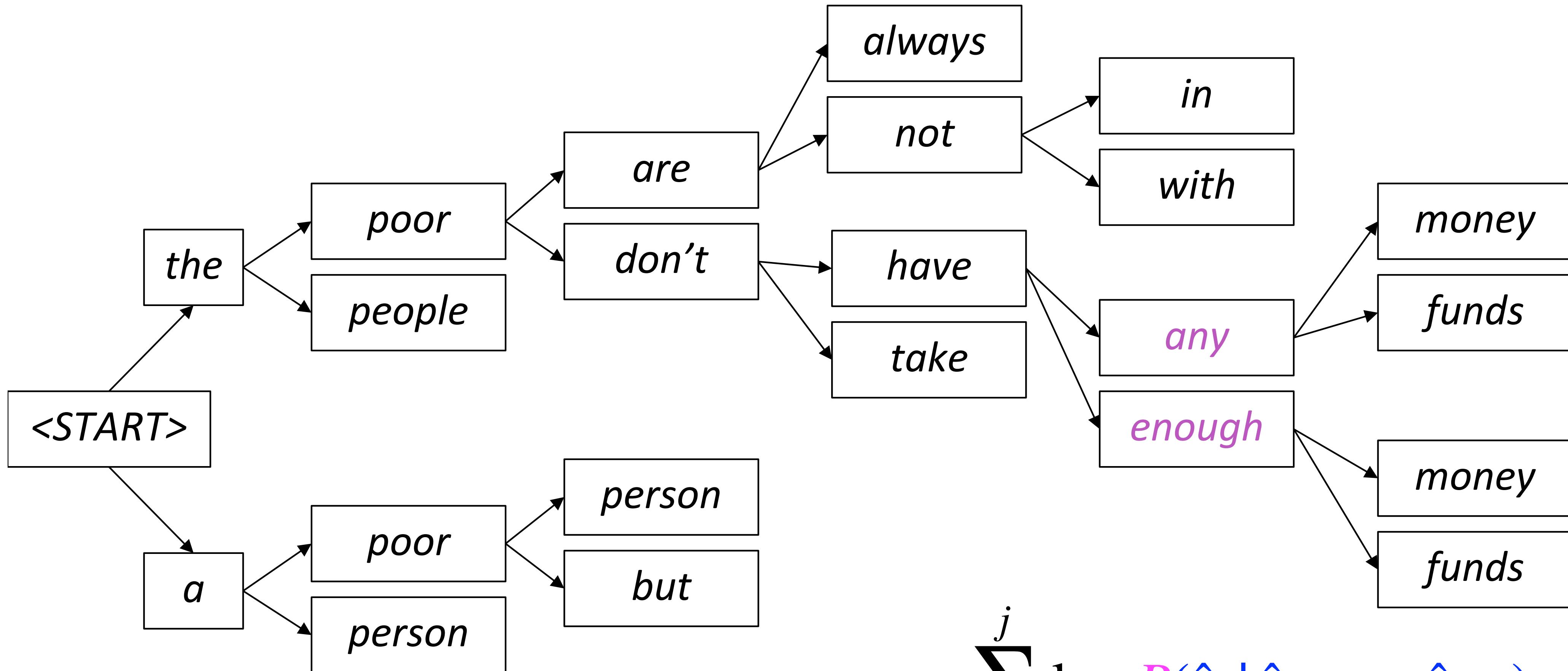


and so on...

$$\sum_{t=1}^j \log P(\hat{y}_t | \hat{y}_1, \dots, \hat{y}_{t-1})$$

Beam Search

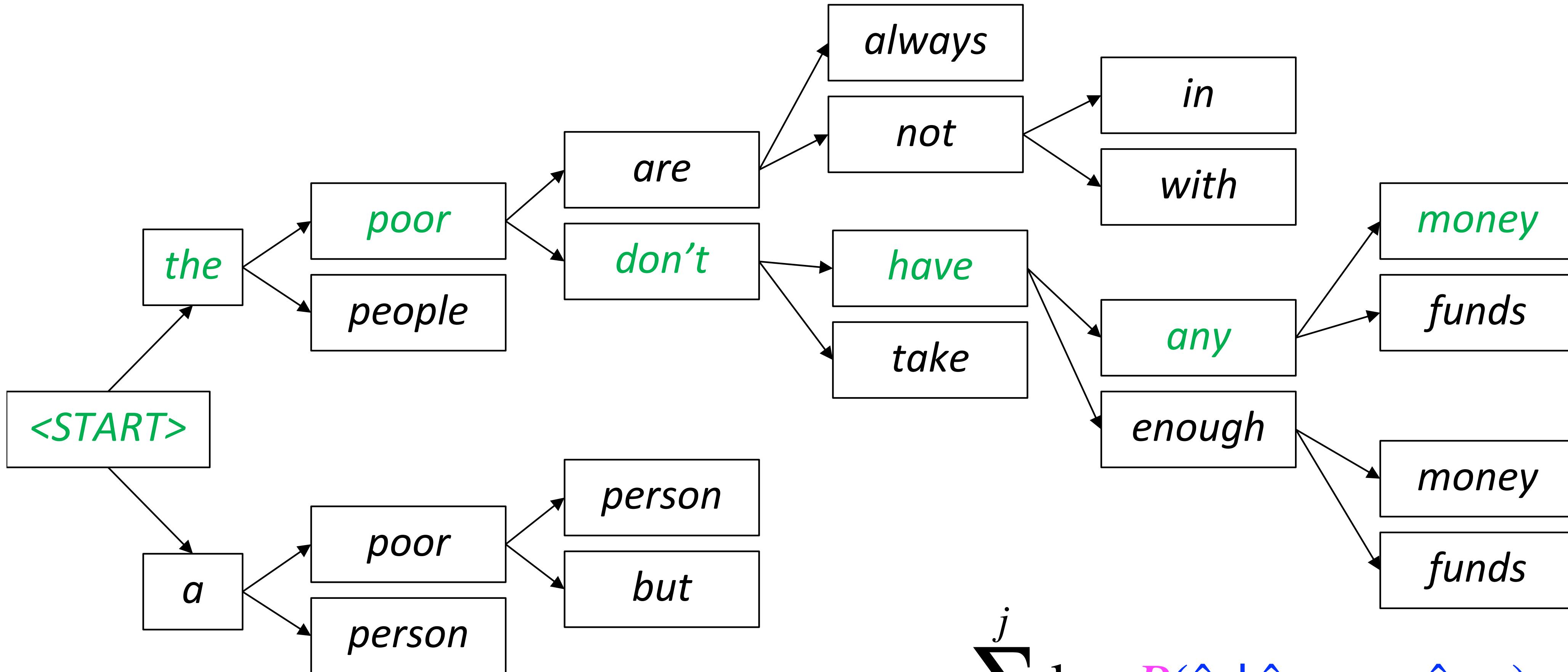
Beam size = 2



$$\sum_{t=1}^j \log P(\hat{y}_t | \hat{y}_1, \dots, \hat{y}_{t-1})$$

Beam Search

Beam size = 2



$$\sum_{t=1}^j \log P(\hat{y}_t | \hat{y}_1, \dots, \hat{y}_{t-1})$$

Beam Search

- Different hypotheses may produce <END> token at different time steps
 - When a hypothesis produces <END>, stop expanding it and place it aside
- Continue beam search until:
 - All b beams (hypotheses) produce <END> OR
 - Hit max decoding limit T
- Select top hypotheses using the *normalized likelihood score*

$$\frac{1}{T} \sum_{t=1}^T \log \textcolor{violet}{P}(\hat{y}_t | \hat{y}_1, \dots, \hat{y}_{t-1}, X)$$

- Otherwise shorter hypotheses have higher scores

**What do you think might happen if we
increase the beam size?**

Effect of beam size

- Small beam size b has similar issues as argmax decoding
 - Outputs that are ungrammatical, unnatural, nonsensical, incorrect
 - $b=1$ is the same as argmax decoding
- Larger beam size b reduces some of these problems
 - Potentially much more computationally expensive
 - Outputs tend to get shorter and more generic