NEURAL TEXT ~~DE~~GENERATION WITH UNLIKELIHOOD TRAINING

Sean Welleck[1,2*] **Ilia Kulikov**[1,2*] **Stephen Roller**[2] **Emily Dinan**[2]

Kyunghyun Cho[1,2,3] **& Jason Weston**[1,2]

Summary

- **Problem:** Standard LM likelihood training and decoding leads to dull and repetitive text

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

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."

- **Bottleneck:** Likelihood objective (token-level model probs are poor)
- **Solution:** New objective, unlikelihood training (assign lower probs to unlikely generations)
- **Result:** Beam search > Sampling method like top-k, nucleus

Language Modeling

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

S = Where are we going

Previous words (Context)

Word being predicted

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

$$\mathcal{L}_{\text{MLE}}(p_\theta, \mathcal{D}) = -\sum_{i=1}^{|\mathcal{D}|} \sum_{t=1}^{|\mathbf{x}^{(i)}|} \log p_\theta(x_t^{(i)}|x_{<t}^{(i)}).$$

Sequence Completion

- **Input:** Prefix (e.g., news story headline, starting of a story)
- **Output:** Continuation (e.g., news story body, rest of a story)

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

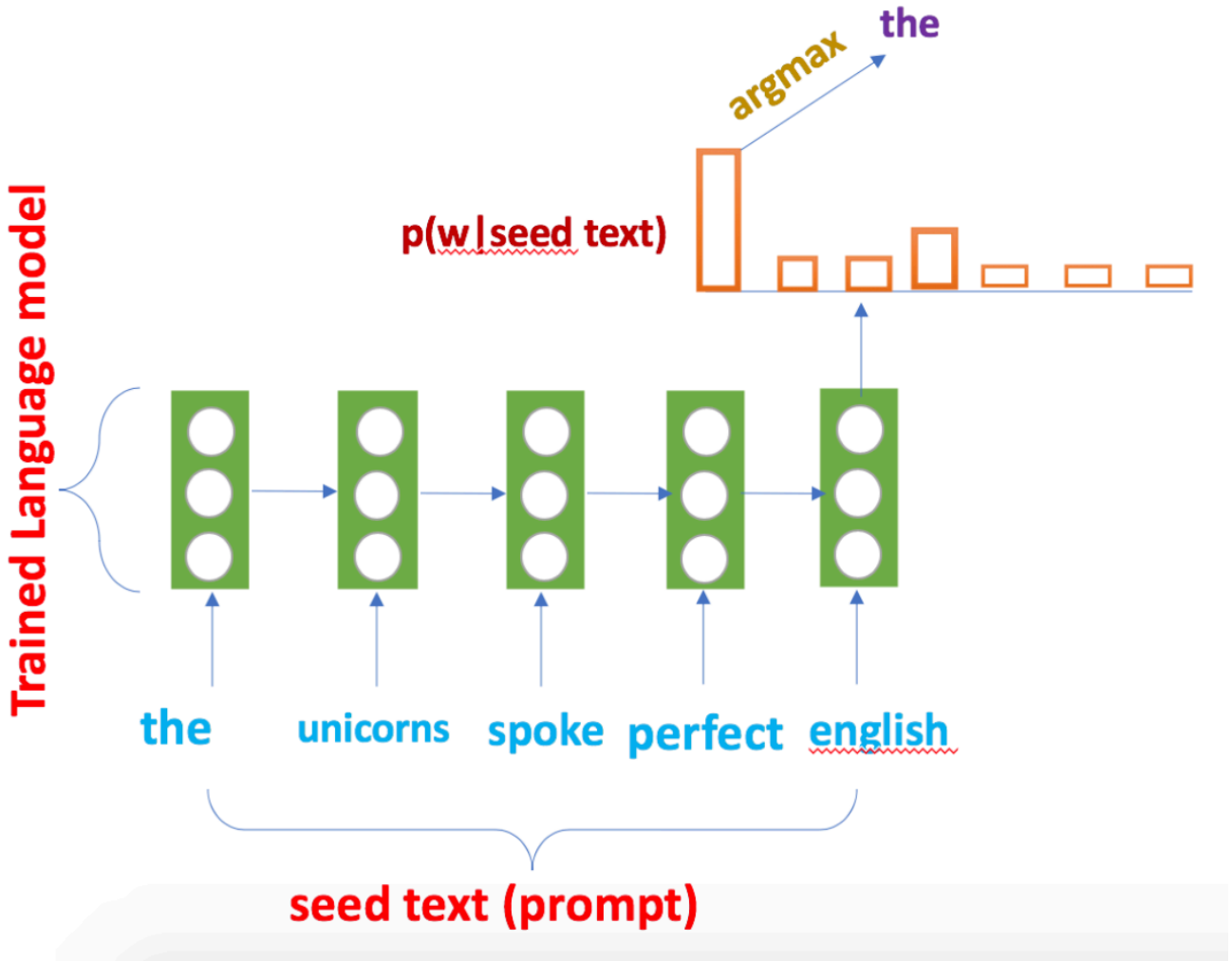
Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

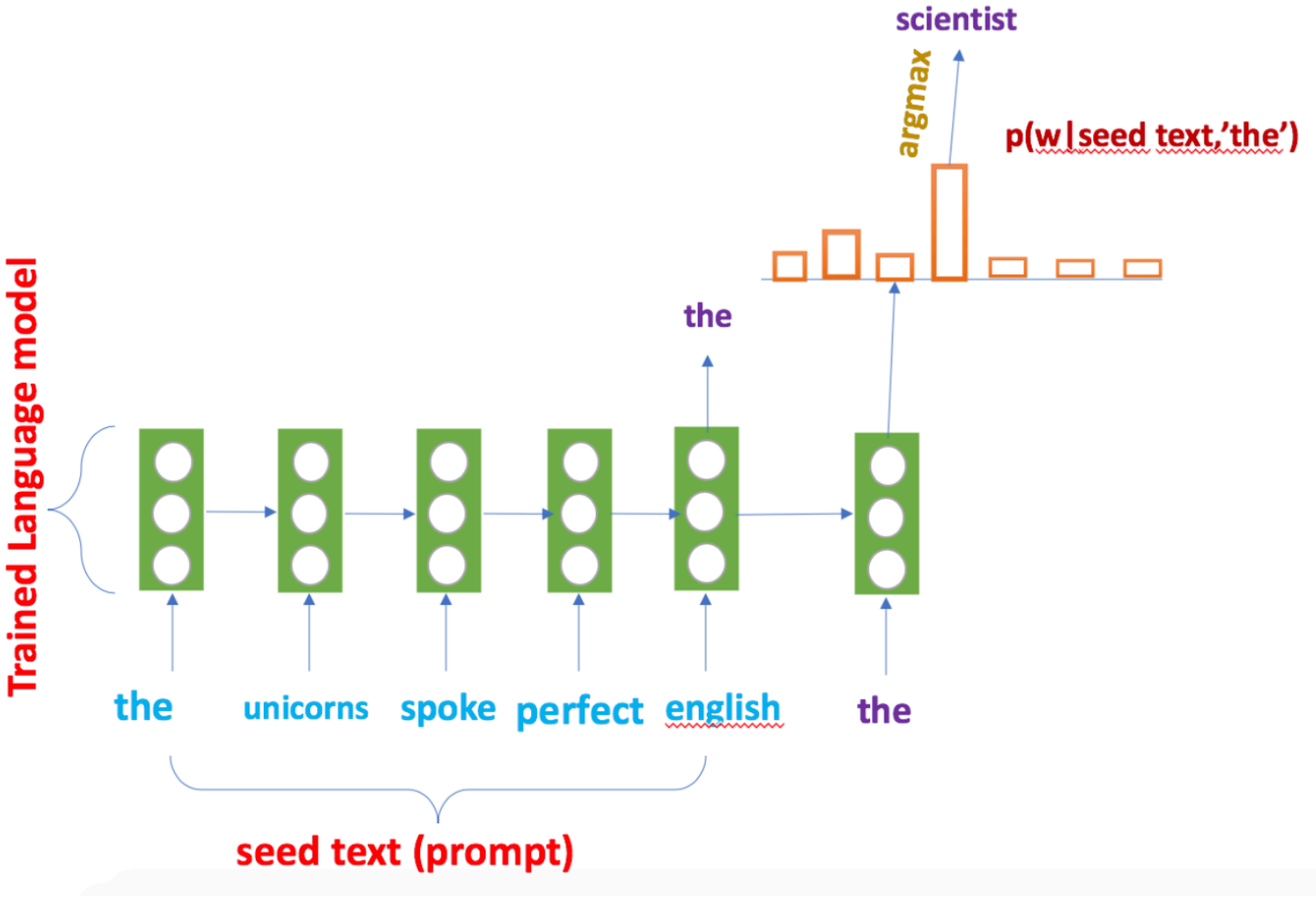
Deterministic decoding techniques

- **Greedy Search**



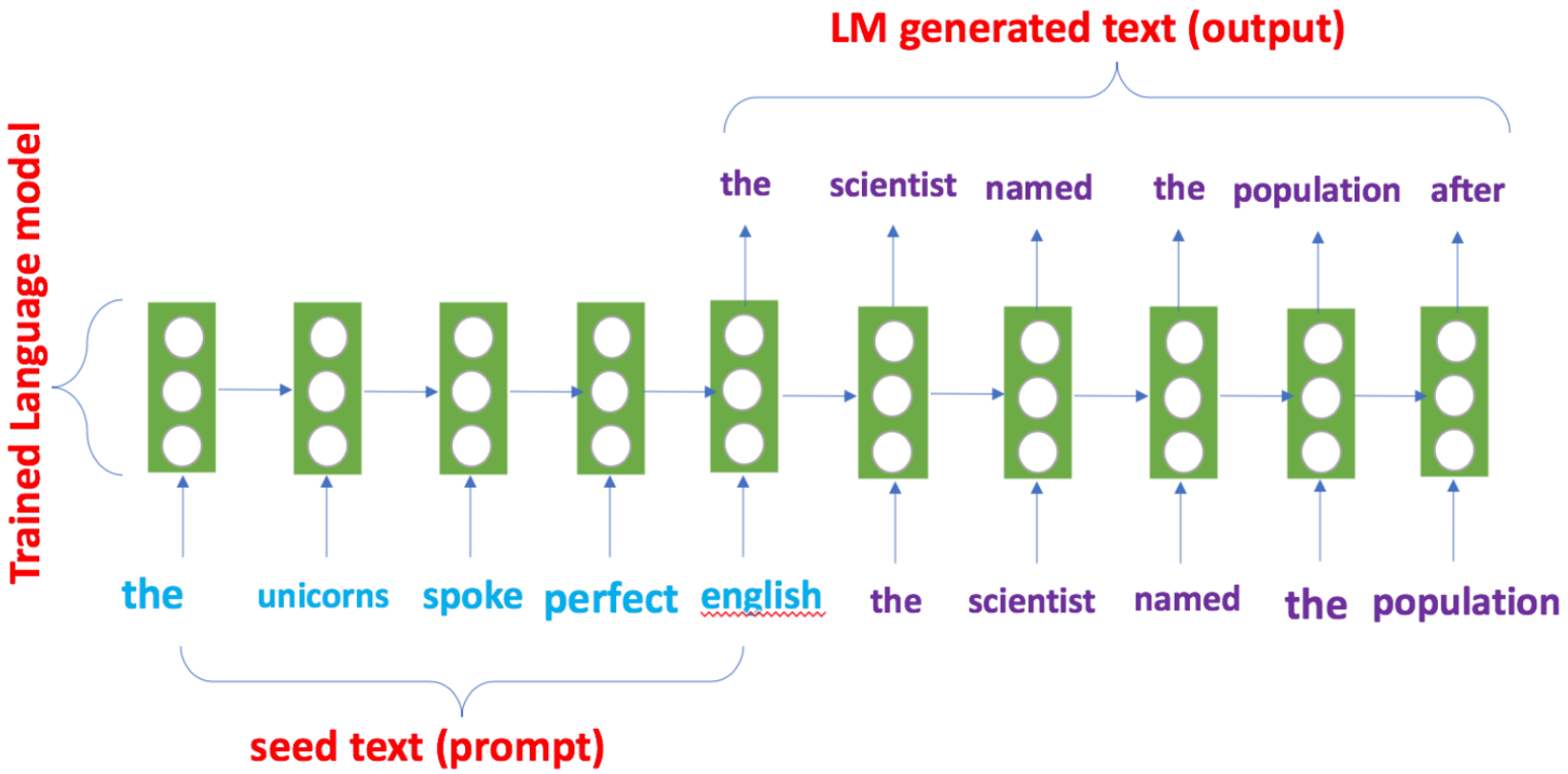
Deterministic decoding techniques

- **Greedy Search**



Deterministic decoding techniques

- **Greedy Search**



Deterministic decoding techniques

- **Beam Search**

Beam search (beam size = 2)

w	p(w|context)
the	0.6
scientist	0.06
named	0.02
population	0.03
a	0.2
french	0.05
researcher	0.04

Beam after 1st timestep = [the (p=0.6), a (p=0.2)]

Deterministic decoding techniques

- **Beam Search**

Beam search (beam size = 2)

w	p(w|context)
the	0.6
~~scientist~~	~~0.06~~
~~named~~	~~0.02~~
~~population~~	~~0.03~~
a	0.2
~~french~~	~~0.05~~
~~researcher~~	~~0.04~~

w	p(w|context, the)
the	0.07
scientist	0.4
named	0.01
population	0.2
a	0.06
french	0.04
researcher	0.02

w	p(w|context, a)
the	0.14
scientist	0.06
named	0.16
population	0.09
a	0.2
french	0.05
researcher	0.3

Beam after 2nd timestep = [the scientist (p=0.4), a researcher (p=0.3)]

Deterministic decoding techniques

- **Beam Search**

Beam search (beam size = 2)

w	p(w|context)
the	0.6
~~scientist~~	~~0.06~~
~~named~~	~~0.02~~
~~population~~	~~0.03~~
a	0.2
~~french~~	~~0.05~~
~~researcher~~	~~0.04~~

Beam after 3rd timestep = [the scientist named (p=0.5), a scientist population (p=0.29)]

w	p(w|context, the)
~~the~~	~~0.07~~
scientist	0.4
~~named~~	~~0.01~~
~~population~~	~~0.2~~
~~a~~	~~0.06~~
~~french~~	~~0.04~~
~~researcher~~	~~0.02~~

w	p(w|context, the, scientist)
the	0.03
scientist	0.07
named	0.5
population	0.29
a	0.01
french	0.08
researcher	0.02

w	p(w|context, a)
~~the~~	~~0.14~~
~~scientist~~	~~0.06~~
~~named~~	~~0.16~~
~~population~~	~~0.09~~
~~a~~	~~0.2~~
~~french~~	~~0.05~~
researcher	0.3

w	p(w|context, the, researcher)
the	0.14
scientist	0.22
named	0.1
population	0.21
a	0.23
french	0.05
researcher	0.05

Degeneration issue with deterministic techniques

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

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."

- a by-product of the model architecture, e.g. the Transformer architecture preferring repeats
- an intrinsic property of human language (humans tend to generate unlikely text)
- a training objective relying on fixed corpora cannot take into account the real goal of using the language

Stochastic Decoding Techniques

- **Top-k sampling**

Top-k sampling (k = 2)

w	p(w|context)
the	0.6
scientist	0.06
named	0.02
population	0.03
a	0.2
french	0.05
researcher	0.04

sort based on p

w	p(w|context)
the	0.6
a	0.2
scientist	0.06
french	0.05
researcher	0.04
population	0.03
named	0.02

Randomly sample a word from top-k words '[the, a]' based on probabilities 'softmax([0.6, 0.2])'

Stochastic Decoding Techniques

- **Top-p (or nucleus) sampling**

Top-p (nucleus) sampling (p = 0.9)

w	p(w|context)
the	0.6
scientist	0.06
named	0.02
population	0.03
a	0.2
french	0.05
researcher	0.04

sort based on p

w	p(w|context)	cum. p.
the	0.6	0.6
a	0.2	0.8
scientist	0.06	0.86
french	0.05	0.91
researcher	0.04	0.95
population	0.03	0.98
named	0.02	1.0

Randomly sample a word from top-p words '[the, a, scientist, french]' based on probabilities 'softmax([0.6, 0.2, 0.06, 0.05])'

Major flaws of likelihood objective

- **Little attention to argmax or top of ranked list of next token probabilities**, instead optimizing the likelihood of the entire distribution
  - Greedy or beam search decoding which rely on the top of the list to generate, are not optimized
  - There is a discrepancy between maximizing the log-probability of a ground-truth token and ensuring the rank of the ground-truth token to be one
- **Not focusing on optimizing sequence generation**, only on producing the next token.
  - During sequence generation, any imperfection in next token prediction leads to error accumulation that is not addressed by likelihood training.

Neural Text Degeneration

- **Repetition**
  - Issue with deterministic decoding
  - Using a Transformer language model trained with likelihood, average percentage of repeated n-grams in model continuations with greedy decoding (43%) far exceeds that of humans (0.5%)
  - Transformer language model predicted next-tokens that appeared in the preceding 128 words 62% of the time, versus 49% in ground-truth text.

$$\Pr\left(\hat{x}_{k+1} = \arg\max p_\theta(x|\mathbf{x}_{1:k}) \in \mathbf{x}_{1:k}\right) > \Pr\left(x_{k+1} \in \mathbf{x}_{1:k}\right)$$

- **Token distribution mismatch**
  - In Transformer language model, the set of next token greedy predictions on a held-out validation set had roughly 40% fewer unique tokens than the ground-truth tokens (11.6k vs. 18.9k)

Proposal

- Two types of update
  - A likelihood update on the true target tokens so that they are assigned high probability
  - **An unlikelihood update on tokens that are otherwise assigned too high a probability**
    - Collect these unlikely token candidates either during next-token prediction or from generated sequences, allowing us to train at both the token and sequence levels
    - Less dull and repetition
    - Improved quality with BS over likelihood + BS or NS

Unlikelihood training objective

$$\mathcal{L}^t_{\text{UL-token}}(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = -\alpha \cdot \underbrace{\sum_{c \in \mathcal{C}^t} \log(1 - p_\theta(c|x_{<t}))}_{\text{unlikelihood}} - \underbrace{\log p_\theta(x_t|x_{<t})}_{\text{likelihood}}.$$

$$\mathcal{C}^t_{\text{prev-context}} = \{x_1, \ldots, x_{t-1}\} \setminus \{x_t\}$$

minimizing the unlikelihood loss with this candidate set makes:

- **incorrect repeating tokens less likely**, as the previous context contains potential repeats
- **frequent tokens less likely**, as these tokens appear often in the previous context

Sequence-level unlikelihood training

- Token-level penalties is limited to prefixes drawn from the training distribution
- Propose sequence-level unlikelihood objective which uses unlikelihood on decoded continuations.

tinuations. That is, given a prefix $(x_1, \ldots, x_k) \sim p_*$, we decode a continuation $(x_{k+1}, \ldots, x_{k+N}) \sim p_\theta(\cdot|x_1, \ldots, x_k)$, construct per-step negative candidate sets $(\mathcal{C}^{k+1}, \ldots, \mathcal{C}^{k+N})$, and define each per-step sequence-level loss for $t \in \{k+1, \ldots, k+N\}$ as:

$$\mathcal{L}^t_{\text{ULS}}(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = -\sum_{c \in \mathcal{C}^t} \log(1 - p_\theta(c|x_{<t})). \quad (7)$$

Intuitively, the negative candidates can identify problematic tokens for the loss to penalize. We choose to penalize repeating n-grams in the continuation:

$$\mathcal{C}^t_{\text{repeat-n}} = \{x_t\} \text{ if } (x_{t-i}, \ldots, x_t, \ldots, x_{t+j}) \in x_{<t-i} \text{ for any } (j-i) = n, i \leq n \leq j, \quad (8)$$

which says that x_t is the (single) negative candidate for step t if it is part of a repeating n-gram[1].

In our experiments we apply this sequence loss in two ways: (i) using it to fine-tune a standard MLE baseline; and (ii) using it to fine-tune an unlikelihood model trained at the token level, $\mathcal{L}_{\text{UL-token}}$. We refer to the former as $\mathcal{L}_{\text{UL-seq}}$ and the latter as $\mathcal{L}_{\text{UL-token+seq}}$. In both cases, fine-tuning is done by equally mixing sequence-level unlikelihood updates (7) and the token-level loss from which it was initially trained (either likelihood updates (1) or token-level unlikelihood updates (4)).

Experiments

- **Model** – Transformer, 16 layer, 8 heads, embed 1024, fully 4096
- **Dataset** – Wikitext-103
- **Training** – prefix length k=50, continuation length N=100, 1.5K updates
- **Evaluation**
  - Repetition

$$\text{rep}/\ell = \frac{1}{|\mathcal{D}|T} \sum_{\mathbf{x} \in \mathcal{D}} \sum_{t=1}^{T} \mathbb{I}\left[\arg\max p_\theta(x|\mathbf{x}_{<t}) \in \mathbf{x}_{t-\ell-1:t-1}\right]$$

A predicted token is called a "single-token repeat" when $\mathbb{I}[\cdot]$ is 1. Some of these single-token repeats also occur in the human-generated sequences, and we thus report a variant which only counts single-token repeats that are additionally not equal to the ground-truth next-token (**wrep/**ℓ).

Experiments

We use the portion of duplicate n-grams (**seq-rep-n**) in a generated sequence to measure sequence-level repetition. That is, for a continuation $\mathbf{x}_{k+1:k+N}$ we compute,

$$\text{seq-rep-n} = 1.0 - \frac{|\text{unique n-grams}(\mathbf{x}_{k+1:k+N})|}{|\text{n-grams}|}, \tag{10}$$

Token Distribution We quantify a model's predicted token distribution using the number of unique tokens. As a token-level metric (**uniq**), we use the number of unique next-token predictions on a validation or test set $\mathcal{D}$, i.e. $|\{\arg\max p(x_t|x_{<t}) \mid x_{<t} \in \mathcal{D}\}|$. As a sequence-level metric (**uniq-seq**) we use the number of unique tokens in continuations of validation or test prefixes (§6).

Language Modeling Quality We use perplexity (**ppl**), and next-token prediction accuracy (**acc**), defined as $\frac{1}{N}|\{\arg\max p(x_t|x_{<t}) = x_t^* \mid x_{<t} \in \mathcal{D}\}|$, with N prefixes $x_{<t}$ and true next tokens x_t^*.

Results

Model	search	seq-rep-4	uniq-seq	ppl	acc	rep	wrep	uniq
$\mathcal{L}_{\text{MLE}}$	greedy	.442	10.8k	**25.64**	**.395**	.627	.352	11.8k
	beam	.523	9.5k					
$\mathcal{L}_{\text{UL-token}}$	greedy	**.283**	**13.2k**	26.91	.390	**.577**	**.311**	**12.7k**
	beam	**.336**	**11.7k**					
$\mathcal{L}_{\text{UL-seq}}$	greedy	.137	13.1k	**25.42**	**.399**	.609	.335	12.8k
	beam	.019	18.3k					
$\mathcal{L}_{\text{UL-token+seq}}$	greedy	**.058**	**15.4k**	26.72	.395	**.559**	**.293**	**13.8k**
	beam	**.013**	**19.1k**					
Human	-	.006	19.8k	-	-	.487	-	19.8k

Table 2: Results for token-level objectives (upper) and sequence-level fine-tuning (lower) according to sequence-level (left) and token-level (right) metrics using the test subset of Wikitext-103.

Results

Search	Model	seq-rep-4	uniq-seq	ppl	acc	rep	wrep	uniq
top-k-3	$\mathcal{L}_{\text{MLE}}$	.0991	14.7k	25.70	.350	.597	.355	12.6k
	$\mathcal{L}_{\text{UL-token}}$	.0491	16.4k	27.02	.344	.539	.306	13.6k
	$\mathcal{L}_{\text{UL-seq}}$	.0068	17.9k	25.11	.353	.581	.341	13.6k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0087	15.2k	26.84	.347	.524	.292	14.6k
top-k-50	$\mathcal{L}_{\text{MLE}}$	.0165	21.9k	25.70	.302	.511	.303	16.1k
	$\mathcal{L}_{\text{UL-token}}$	.006	23.5k	27.02	.286	.440	.247	17.8k
	$\mathcal{L}_{\text{UL-seq}}$	.0005	25.7k	25.11	.291	.497	.291	17.3k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0009	23.7k	26.84	.289	.430	.238	18.8k
top-p-0.3	$\mathcal{L}_{\text{MLE}}$	.273	13.6k	25.70	.264	.339	.154	12.6k
	$\mathcal{L}_{\text{UL-token}}$	.101	16.5k	27.02	.247	.290	.121	13.9k
	$\mathcal{L}_{\text{UL-seq}}$	.0033	20.8k	25.11	.266	.327	.145	13.6k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0041	19.1k	26.84	.250	.284	.116	14.9k
top-p-0.9	$\mathcal{L}_{\text{MLE}}$	.0154	26.9k	25.70	.288	.462	.263	18.6k
	$\mathcal{L}_{\text{UL-token}}$	.004	30.2k	27.02	.266	.381	.202	22.3k
	$\mathcal{L}_{\text{UL-seq}}$	.0003	34.7k	25.11	.290	.450	.254	19.6k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0007	32.4k	26.84	.269	.376	.198	22.7k
Human	-	.006	19.8k	-	-	.487	-	19.8k

Table 6: Stochastic decoding results according to sequence-level (left) and token-level (right) metrics using the test subset of Wikitext-103.

Results

Which writer is from Wikipedia?

You are currently at comparison 1 / 5

You will read an excerpt from Wikipedia and two possible continuations. Judge whether Writer 1 on the left or Writer 2 on the right wrote the next few sentences in Wikipedia.

You should prefer the continuation which is **more likely** based on content and style. You should also provide a very brief justification. Failure to do so could result in your hits being rejected.

You may need to scroll down to see the full excerpts.

You will do this for 5 pairs of conversations. After completing each judgement, use the [NEXT] button (which will appear below after you finish your judgement).

Beginning of excerpt:

…, which had a major negative effect on the state 's large merchant fleet. Gore was in 1808 elected to the Massachusetts House of Representatives, where he successfully led Federalist efforts to ensure the selection of a Federalist slate of presidential electors. He also spearheaded actions to…

Writers' continuations:

…prevent the American Revolutionary War from being fought by the British, and to prevent the British from using the country to launch a war against the British. Gore 's actions in the House of Representatives were a major turning point in his political career. He was elected to the House of Representatives in 1811, and served until his death in 1815.

Early life and education

Gore was born in Boston, Massachusetts, on February 22, 1798, the son of Benjamin Gore and his …

… drive Senator John Quincy Adams from the Federalist Party over his support of Thomas Jefferson 's foreign policy. The legislature elected Adams ' successor nine months early, and gave Adams sufficiently distasteful instructions that he resigned the post and joined with the Republicans.

Governor

Gore led the Federalists to victory in 1809 against Sullivan 's successor, Levi Lincoln, Sr., who had taken over as acting governor upon Sullivan 's death late in 1808. During Gore 's term the principal domestic issue occupying state politics…

Which writer's work is more likely to be in Wikipedia?

○ Writer 1 's is more likely ○ Writer 2 's is more likely

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

Results

		seq-rep-4
Prefix	*... Lyrically the song has excerpts of different languages including French , Spanish*	
$\mathcal{L}_{\text{MLE}}$	, Italian , Spanish , Italian , Spanish , Italian , Spanish , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese	0.744
$\mathcal{L}_{\text{UL-token+seq}}$	, and German . In the first verse , the protagonist sings about being a " girl who 's been in love with someone else " , while the second verse describes the relationship between the protagonist and her lover . In the third verse , the protagonist sings	0.063
Prefix	*... starboard engines and was going to crash . " We 're going in ,"*	
$\mathcal{L}_{\text{MLE}}$	he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to	0.787
$\mathcal{L}_{\text{UL-token+seq}}$	Hood said . " I 'm going to make sure we 're going to get back to the water . " The order to abandon ship was given by Admiral Beatty , who ordered the remaining two battlecruisers to turn away . At 18 : 25 , Hood turned his	0.000
Prefix	*... career - high 27 points on 8 - for - 11 shooting with three rebounds*	
$\mathcal{L}_{\text{MLE}}$	and two assists . On January 3 , 2012 , he was named to the 2012 13 All - Atlantic 10 first team . On February 3 , 2012 , he was named to the Atlantic 10 first team . On February 5 , 2012 , he was named	0.277
$\mathcal{L}_{\text{UL-token+seq}}$	and a career - high 7 assists against the Minnesota Timberwolves . On February 3 , 2012 , he was named to the 2012 All - NBA First Team . On March 7 , 2012 , he was named one of five finalists for the Naismith Award , which is	0.064

Table 1: Example greedy completions showing representative examples of the MLE model's degenerate single-token repetition (top), phrase-level repetition (middle), and 'structural' repetition (bottom), as well as the proposed method's ability to fix these degenerate behaviors.

Results

Winner		Loser	Crowdworkers Win rate	Experts Win rate
$\mathcal{L}_{\text{UL-token}}$		$\mathcal{L}_{\text{MLE}}$ baseline	57%	
$\mathcal{L}_{\text{UL-seq}}$		$\mathcal{L}_{\text{MLE}}$ baseline	*71%	
$\mathcal{L}_{\text{UL-token+seq}}$	*beats*	$\mathcal{L}_{\text{MLE}}$ baseline	*82%	
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{UL-token}}$	*75%	
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{UL-seq}}$	59%	
$\mathcal{L}_{\text{UL-token+seq}}$	*beats*	$\mathcal{L}_{\text{MLE}}$ Nucleus sampling ($p = 0.9$)	59%	*83%
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{MLE}}$ Beam blocking (4-gram)	60%	*74%

Table 3: **Human eval results**. * denotes statistical significance (2-sided binomial test, $p < .05$).