

Neural Text (*De*)generation

Jan Buys

University of Cape Town

*Joint work with Ari Holtzman, Li Du, Maxwell Forbes and
Yejin Choi*





ROALD DAHL
**THE GREAT
AUTOMATIC
GRAMMATIZATOR**
& OTHER STORIES



The Great Grammatizator?

Prompt: *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.*

The Great Grammatizator?



Machine-written continuation: 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.

OpenAI GPT-2 (February 2019)

Natural Language Generation in Practice



Machine Translation

Google Translate

DETECT LANGUAGE **AFRIKAANS** ENGLISH SPANISH ▾ ↕

Regter Raymond Zondo se kommissie van ondersoek na staatskaping ✕
sit in 'n onmerkwaardige moderne kantoorgebou in Empireweg in
Johannesburg. Dit is hier op die vierde vloer waar getuies die
reputasies van gewese helde stukkie vir stukkie vernietig en Suid-
Afrika se geskiedenis herskryf.

🔊 286/5000 🗑️ ▾

Machine Translation

Google Translate



ENGLISH

AFRIKAANS

SPANISH



Judge Raymond Zondo's commission of inquiry into state capture sits in an unimaginable modern office building in Empire Road in Johannesburg. It is here on the fourth floor where witnesses destroy the reputation of former heroes piece by piece and rewrite South Africa's history.



Automatic summarization

michelle pfeiffer is the latest hollywood star preparing to hit the small screen . the oscar nominated star known for her roles in iconic films such as scarface , dangerous liaisons and the age of innocence , has teamed up with katie couric to pitch a new television comedy about a morning news program . also involved in the project , in which pfeiffer is attached to star , is diane english , the creator of murphy brown . scroll down for video michelle pfeiffer (left) is set to star in a new television comedy about a morning news program produced by katie couric (right) the series was created by diane english (above with candice bergen) , who was behind the show murphy brown , about a female news anchor according to variety , pfeiffer ' s role will be that of a morning news anchor , making it very similar to the real life role couric played as co - host of today for 15 years . couric will serve as an executive producer and help ' ensure the series strikes realistic notes . ' the creator behind the project , english , was previously the brains behind brown , the show starring candice bergen that centered around a female news anchor and ran for ten seasons , winning 18 emmys . english would also serve as a writer , producer and showrunner on the program . the ladies are currently in talks with hbo , showtime , amc , netflix and amazon to pick up the program . couric will serve as an executive producer , drawing on her experience as an anchor on today for 15 years pfeiffer would be the one of the biggest stars yet to move to television ...



michelle *pfeiffer* is set to star in a new tv comedy about a morning news program . couric will serve as an executive producer and showrunner on the project . the series was created by diane english , the creator of murphy brown . pfeiffer is one of the biggest stars.

(Celikyilmaz et al., 2018)

Natural Language Processing



Large Pre-trained Language Models: ELMo, GPT, BERT, RoBERTa, etc

Fake news?

Machine-written continuation: ...The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of

OpenAI's GPT-2 Transformer Language Model (released version with 117M parameters) decoded with beam search.

Chatbots

[Human:] Hello there, I like rock climbing do you have any hobbies?

[Bot:] i like nature. do you? i like nature.

[Human:] I do love nature, especially when going on walks

[Bot:] what is your favorite rock? mine is rock.

[Human:] Well I like rock music, does that count?

[Bot:] i think it does. what do you do for fun?

[Human:] Well I work at a restaurant that I love, and rock climbing

[Bot:] that sounds fun! i'm into nature.

[Human:] Cool, so am I

[Bot:] do you have any pets? do you have pets?

Talk overview

- What is the best method to decode text from neural language models?
- How can we improve the model to guide it towards generating high-quality text?

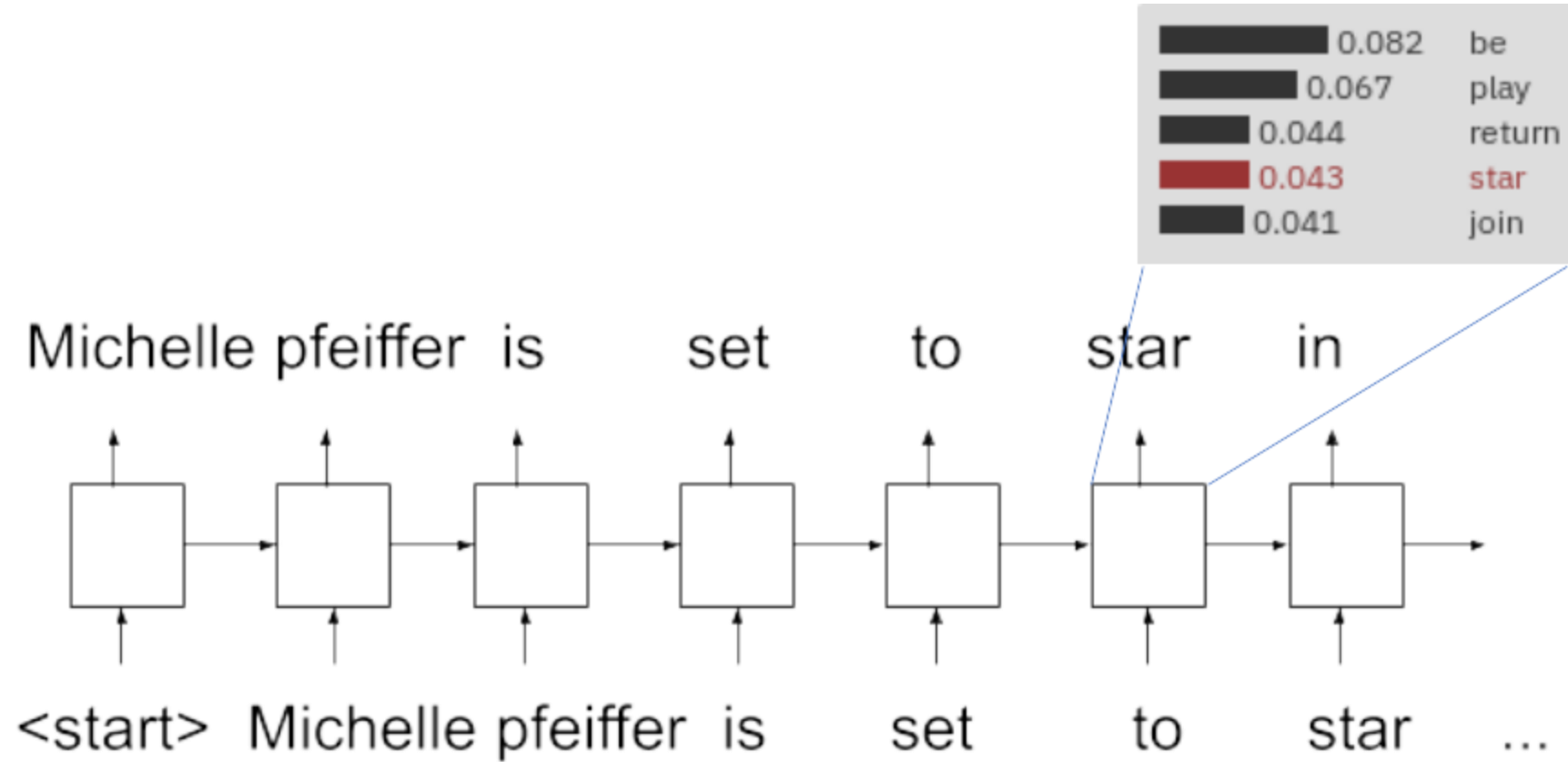
Language Modelling

P (Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.)

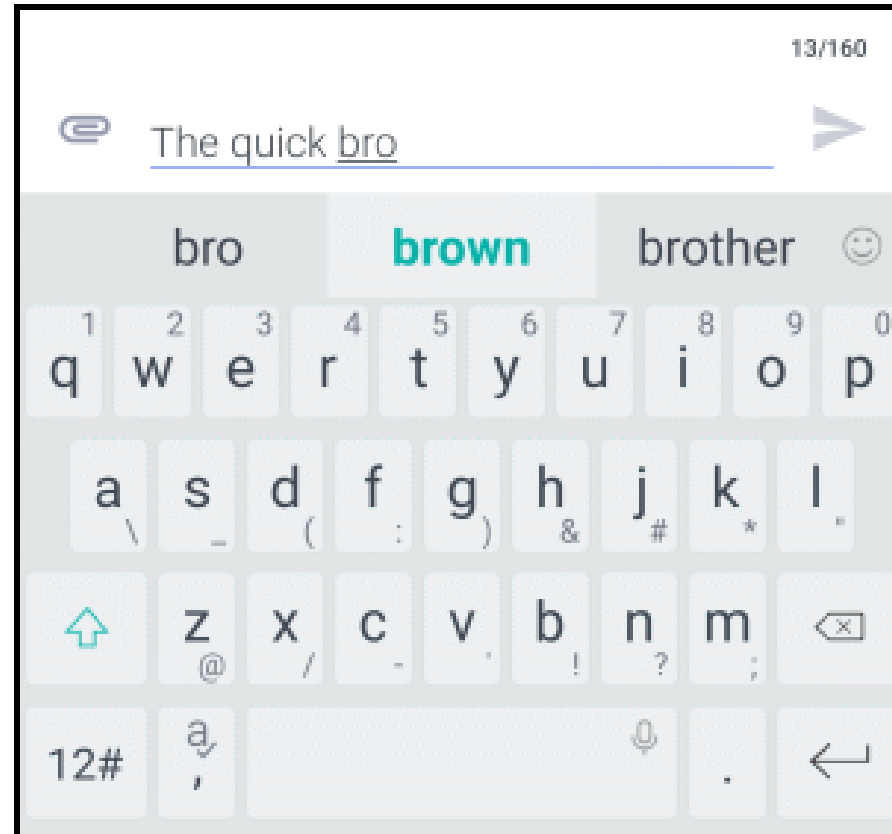
Language Modelling

$$P(x_{1:m+n}) = \prod_{i=1}^{m+n} P(x_i | x_1 \dots x_{i-1})$$

Language Modelling



Language Modelling



Text Generation from Language Models



WebText

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



Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Open-ended Generation with Neural Language Models

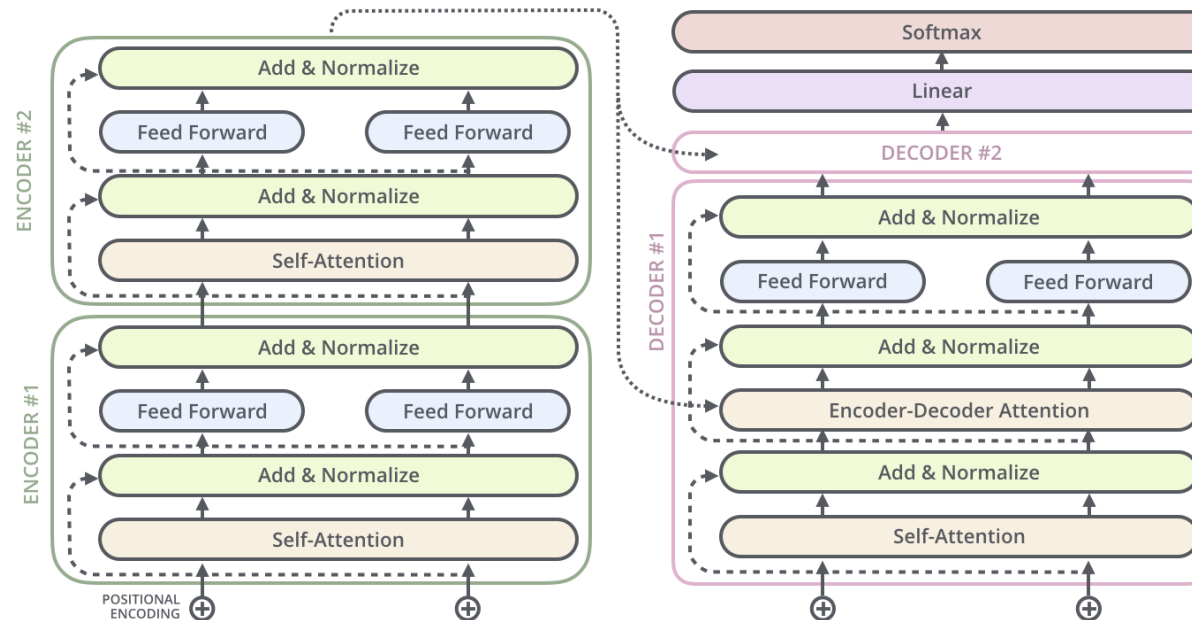
- Generate **word sequence** \mathbf{x} given **context** \mathcal{Y}

$$P(\mathbf{x} \mid \mathcal{Y})$$

$$= \prod_i P(x_i \mid \mathbf{x}_1^{i-1}, \mathcal{Y})$$

Text Generation from Language Models

- GPT-2 Large, trained on WebText (40GB)
- Generate 200 tokens, conditioned on document's first paragraph



Decoding Text from Language Models

- Maximization-based decoding
- Sampling-based decoding
 - Sampling with temperature
- Truncated sampling
 - Top-k sampling
 - **Nucleus Sampling**

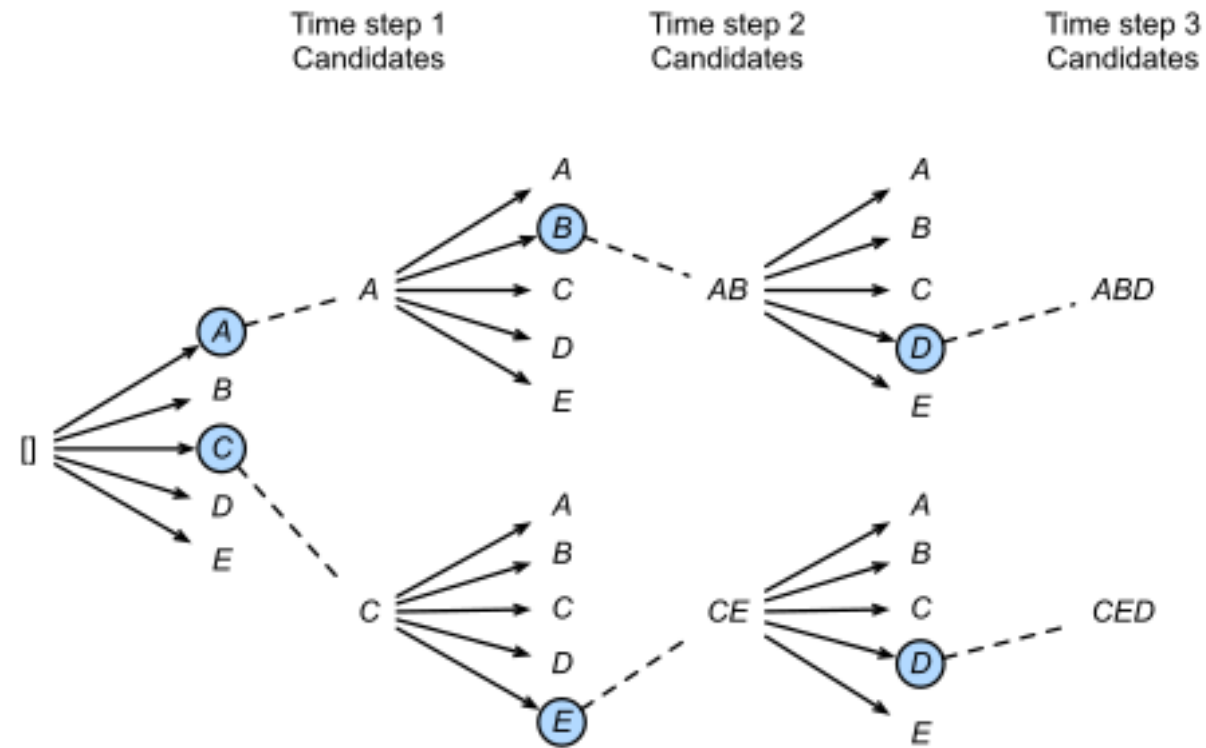
Maximization-based Decoding

- Greedy decoding: Choose the highest scoring word at each time step

$$x_i = \operatorname{argmax} P(x_i | x_1 \dots x_{i-1})$$

Maximization-based Decoding

- Beam search



Maximization-based Decoding

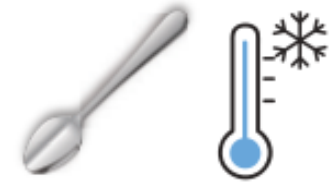
- **Context:** An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.
- **Generation:** The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing 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.

Sampling-based Decoding



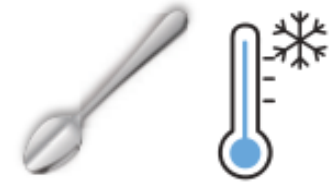
- **Generation:** The Australian Food Safety Authority has warned Australia's beaches may be **revitalised** this year because healthy **seabirds and seals** have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by **the Holden CS118 and Adelaide Airport 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 with Temperature



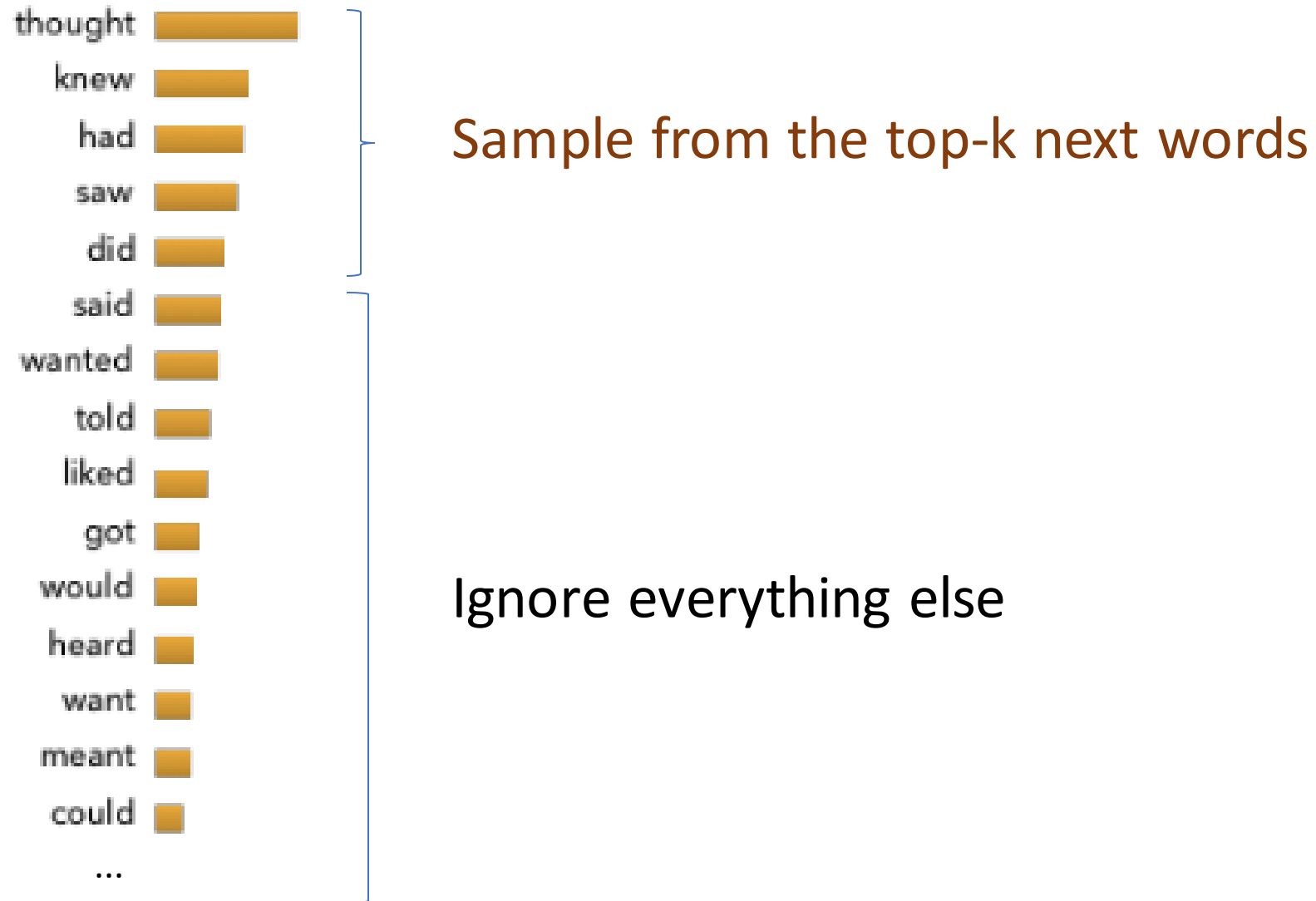
$$p(x = V_l | x_{1:i-1}) = \frac{\exp(u_l/t)}{\sum_{l'} \exp(u_{l'}/t)}.$$

Sampling with Temperature



- **Generation:** 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.**

Truncated sampling: Top-k Sampling



Truncated sampling: Top-k Sampling

- **Top 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](http://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.

Truncated sampling: Nucleus Sampling



Sample from the words with the top-p probability mass

Ignore everything else

Truncated sampling: Nucleus Sampling

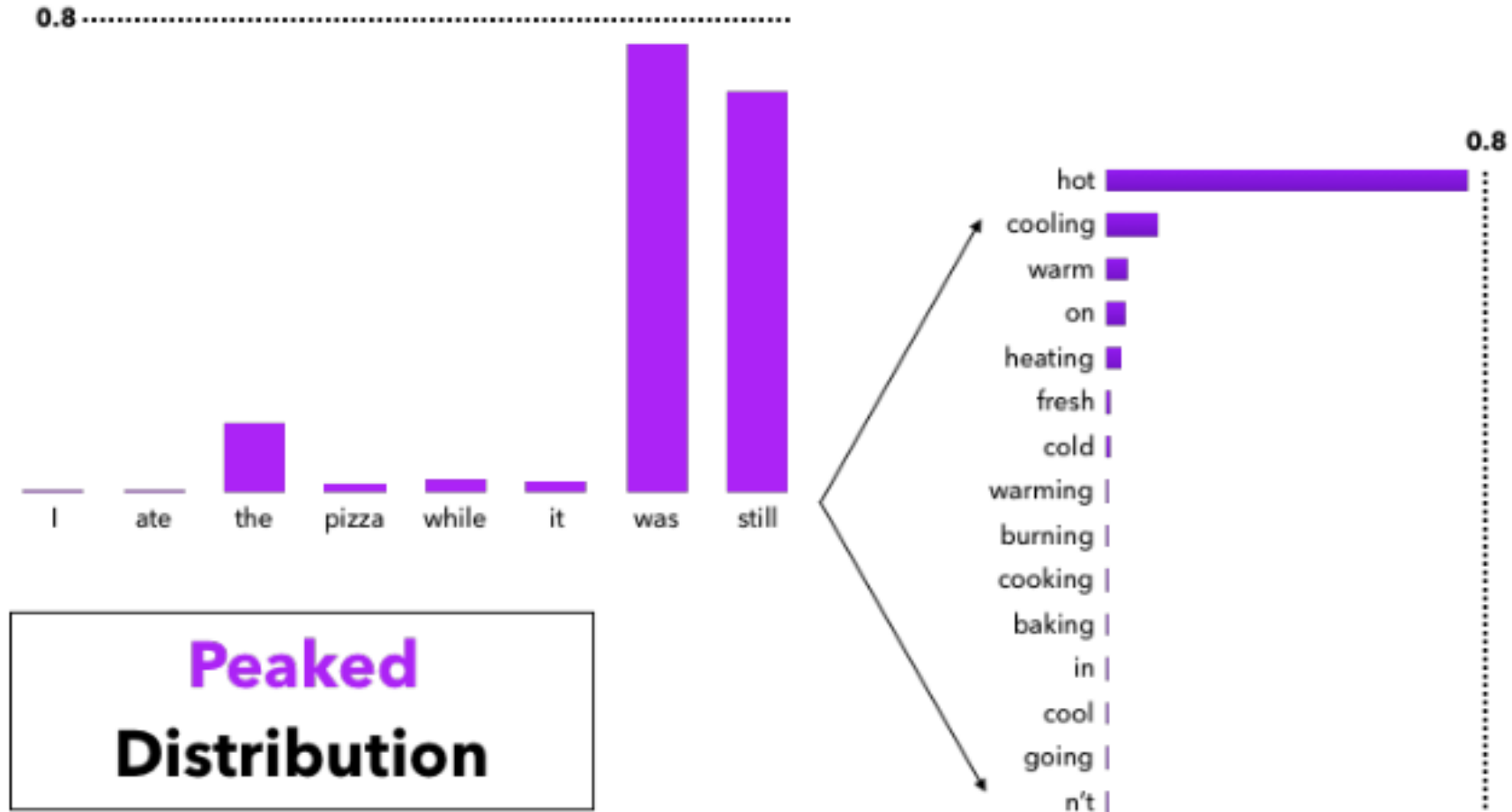
- Define the top-p vocabulary $V^{(p)} \subset V$ as the smallest set such that

$$\underbrace{\sum_{x \in V^{(p)}} P(x|x_{1:i-1})}_{p'} \geq p.$$

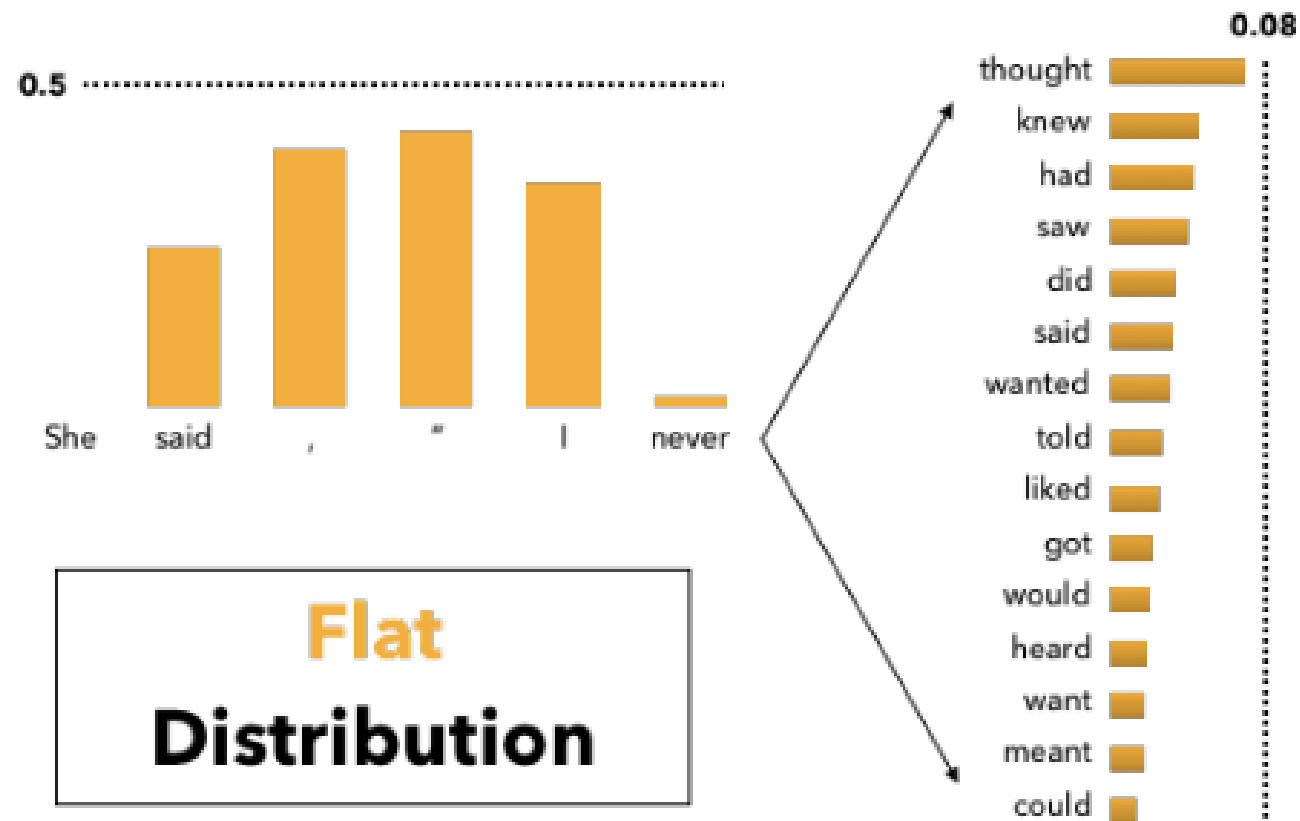
- The distribution sampled from is defined as

$$P'(x|x_{1:i-1}) = \begin{cases} P(x|x_{1:i-1})/p' & \text{if } x \in V^{(k)} \\ 0 & \text{otherwise.} \end{cases}$$

Truncated sampling: Nucleus Sampling



Truncated sampling: Nucleus Sampling



Truncated sampling: Nucleus Sampling

- **Top 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**.

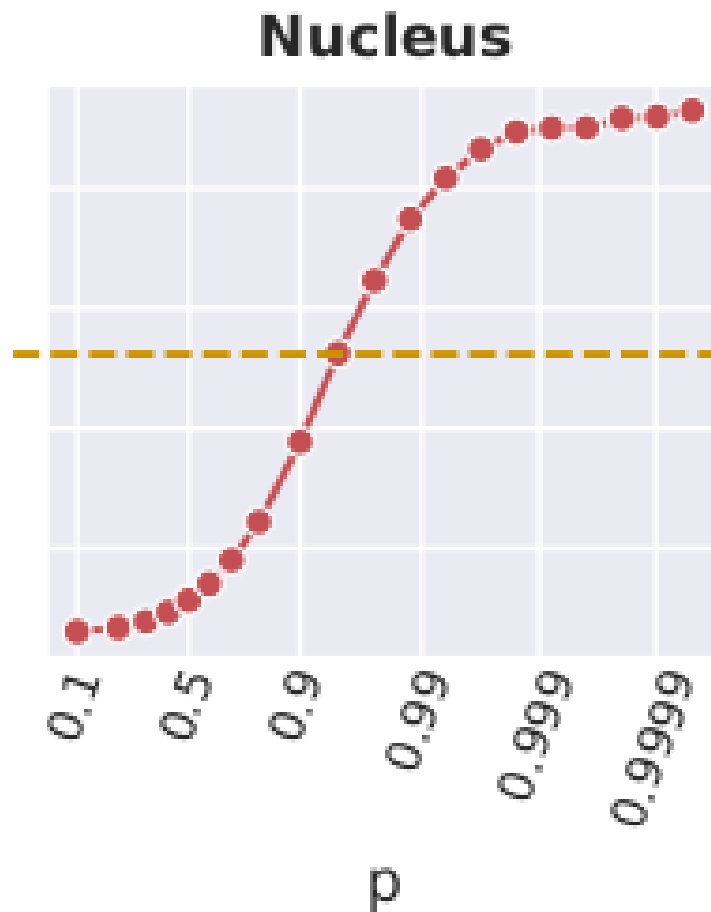
Evaluation

- Quality: Perplexity
- Diversity:
 - Self-BLEU
 - Vocabulary usage distribution
 - Repetition
- Human + machine evaluation

Evaluation: Perplexity of Generated Text

| Method | Perplexity |
|---------------------------|--------------|
| Human | <u>13.08</u> |
| Greedy | 1.50 |
| Pure sampling | 22.73 |
| Sampling, temperature 0.9 | 10.25 |
| Top-k sampling, k=640 | 13.82 |
| Nucleus sampling, p=0.95 | 13.13 |

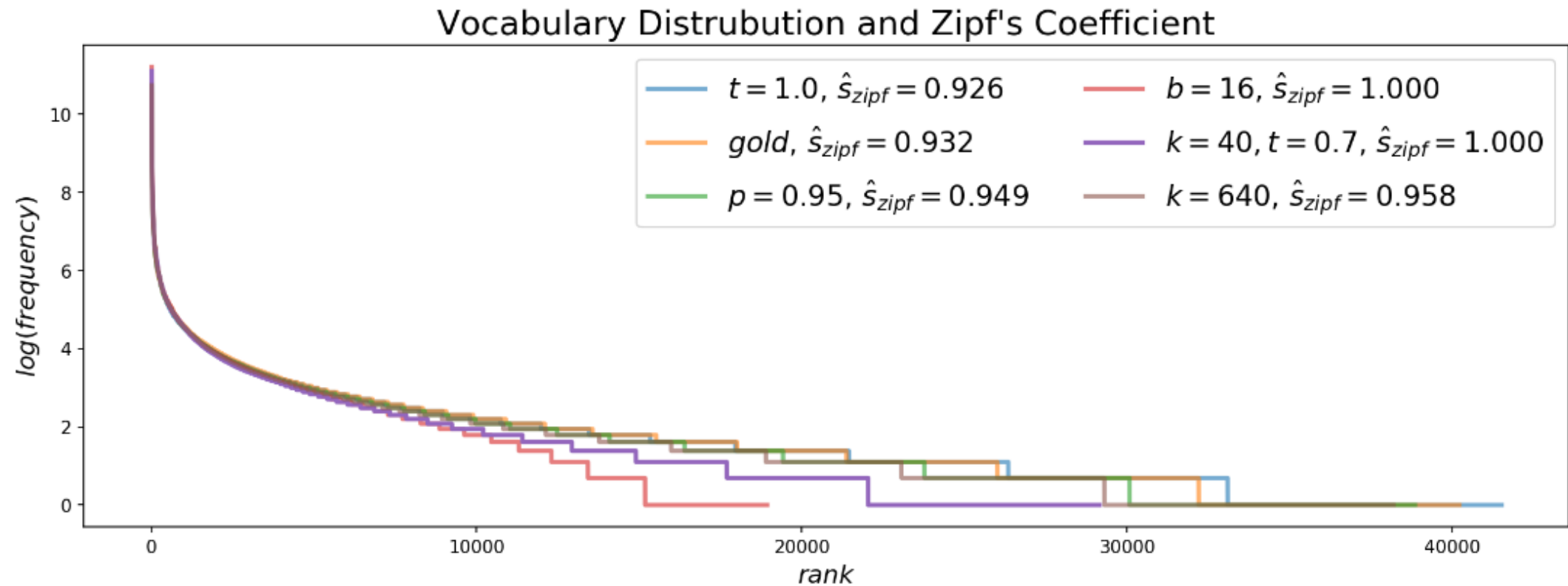
Evaluation: Perplexity of Generated Text



Diversity Evaluation: Self-BLEU

| Method | Self-BLEU-4 |
|---------------------------|-------------|
| Human | <u>0.30</u> |
| Greedy | 0.50 |
| Pure sampling | 0.28 |
| Sampling, temperature 0.9 | 0.35 |
| Top-k sampling, k=640 | 0.32 |
| Nucleus sampling, p=0.95 | 0.32 |

Diversity Evaluation: Vocabulary distribution



Diversity Evaluation: Repetition

| Method | Repetition loop % |
|---------------------------|-------------------|
| Human | <u>0.18</u> |
| Greedy | 74.20 |
| Pure sampling | 0.14 |
| Sampling, temperature 0.9 | 0.54 |
| Top-k sampling, k=640 | 0.14 |
| Nucleus sampling, p=0.95 | 0.20 |

Evaluation

- Quality: Perplexity
- Diversity:
 - Self-BLEU
 - Vocabulary usage distribution
 - Repetition
- Human + machine evaluation

Evaluation: Human + Automatic

- Human Unified with Statistical Evaluation (HUSE) - Hashimoto et al. (2019)
- Train a classifier to discriminate between human and machine text using 2 features:
 - Language Modelling score (diversity)
 - Human judgement score (coherence)
- Truncated sampling: Interpolate with original distribution

Evaluation: Human + Automatic

| Method | HUSE score |
|---------------------------|-------------|
| Pure sampling | 0.67 |
| Sampling, temperature 0.9 | 0.79 |
| Top-k sampling, k=40 | 0.19 |
| Top-k sampling, k=640 | 0.94 |
| Nucleus sampling, p=0.95 | 0.97 |

Summary so far

- Language model's probability distributions are systematically wrong in ways which adversely affects language generation quality
- Truncated sampling mitigates this problem
- Nucleus sampling is a principled and well-behaved heuristic for improving generation quality

Open questions

- Have we solved the problem of neural text degeneration?
- Can all remaining issues be resolved with:
 - Larger models?
 - More training data?
 - More complex network architectures?
- Is language modelling alone enough to generate text?

Learning to Write

- How can we teach neural networks to write?
- What is the appropriate objective function for writing?

Neural Generation can be Brittle

"I'm so sorry for your loss," he said. "I'm so sorry." "I'm sorry," she said. "I'm sorry." "Don't be."

Adaptive LSTM Language Model trained on Toronto BookCorpus (**980 million words**) decoded with stochastic top-k beam search.

Neural Generation can be Brittle

"I'm so sorry for your loss," he said. "I'm so sorry." "I'm sorry," she said. "I'm sorry." "Don't be."

Repetition

Neural Generation can be Brittle

He was dressed in a dark black suit and white shirt. He wore jeans and a black tee shirt. He was dressed in a simple white shirt and black pants. His hair was cut short, which he wore to the occasion. He was wearing a navy blue shirt and a pair of dark blue jeans.

Contradiction

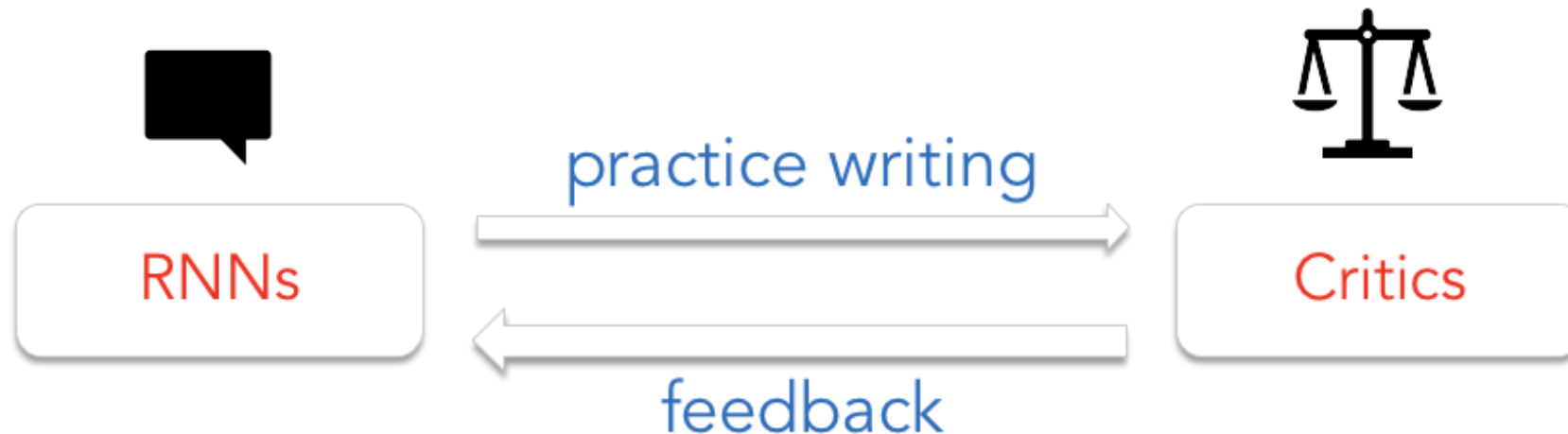
Neural Generation can be Brittle

What the hell is that supposed to mean? I'm not sure, but I'm pretty sure they are. I mean, what are you doing here? The question is, why are you here? I don't know.

Generic, uninformative

Learning to Write with Cooperative Discriminators

- Let a committee of critics judge the LM-generated text – and use their feedback to teach the model to learn to write better



Grice's Maxims

- Principles underlying conversation or communication
 1. Be truthful
 2. Be informative
 3. Be relevant
 4. Be clear

Grice's Maxims

1. Be truthful
2. Be informative
3. Be relevant
4. Be clear

Violations:

Contradiction

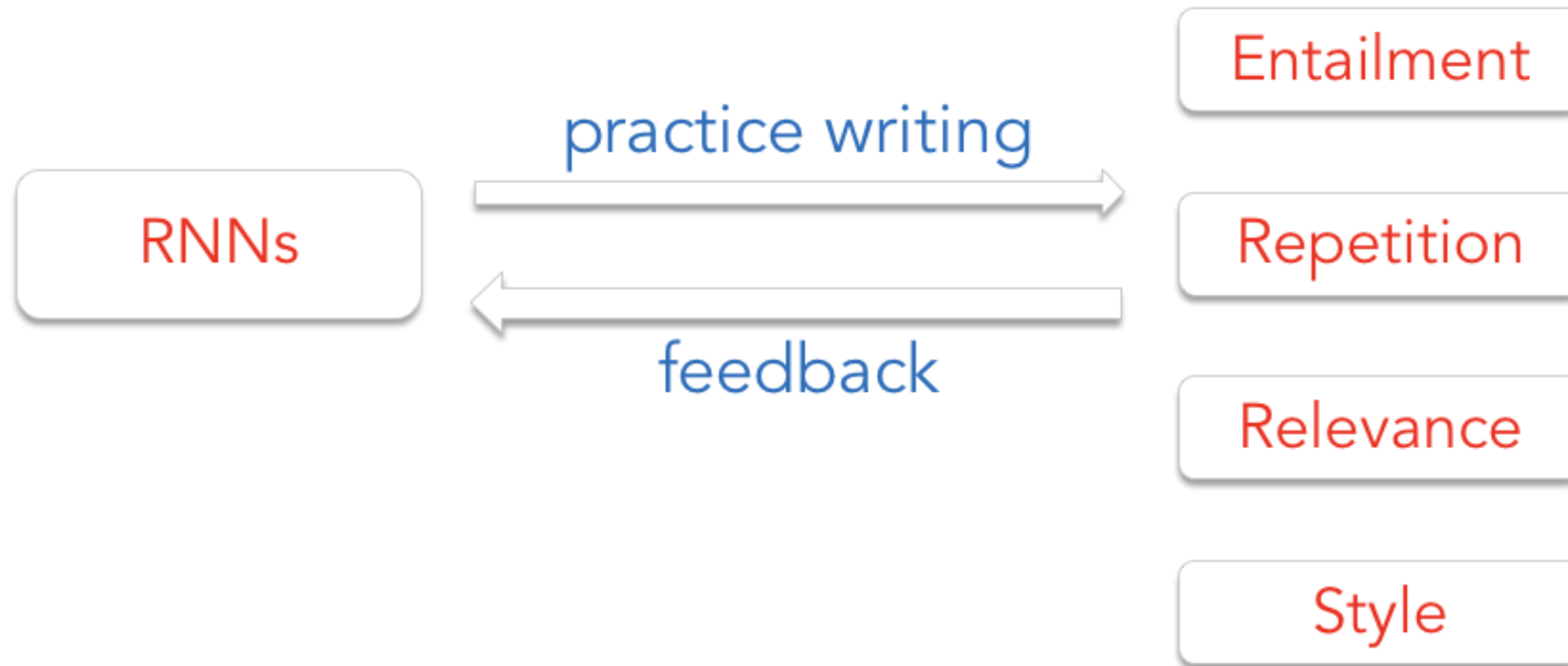
Redundancy, Repetition

Digression

Vagueness, obscurity

Discriminators inspired by Grice's Maxims

Truthful, Informative, Relevant, Clear



Entailment Module

Be Truthful

He was dressed in a dark black suit and white shirt.



He wore jeans and a black tee shirt.

Entailment Module

Be Informative

He was dressed in a dark black suit and white shirt.



He was dressed in a white shirt and black pants.

Entailment Module

He was dressed in a white shirt.



He was wearing a pair of dark blue jeans.

Entailment Module

- Both entailment and contradiction should be penalized
- Train classifier on large entailment datasets

$$s_{\text{entail}}(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{a} \in S(\mathbf{x}) \cup S_{\text{init}}(\mathbf{y})} \text{neutral}(\mathbf{a}, S_{\text{last}}(\mathbf{y}))$$

Repetition Module

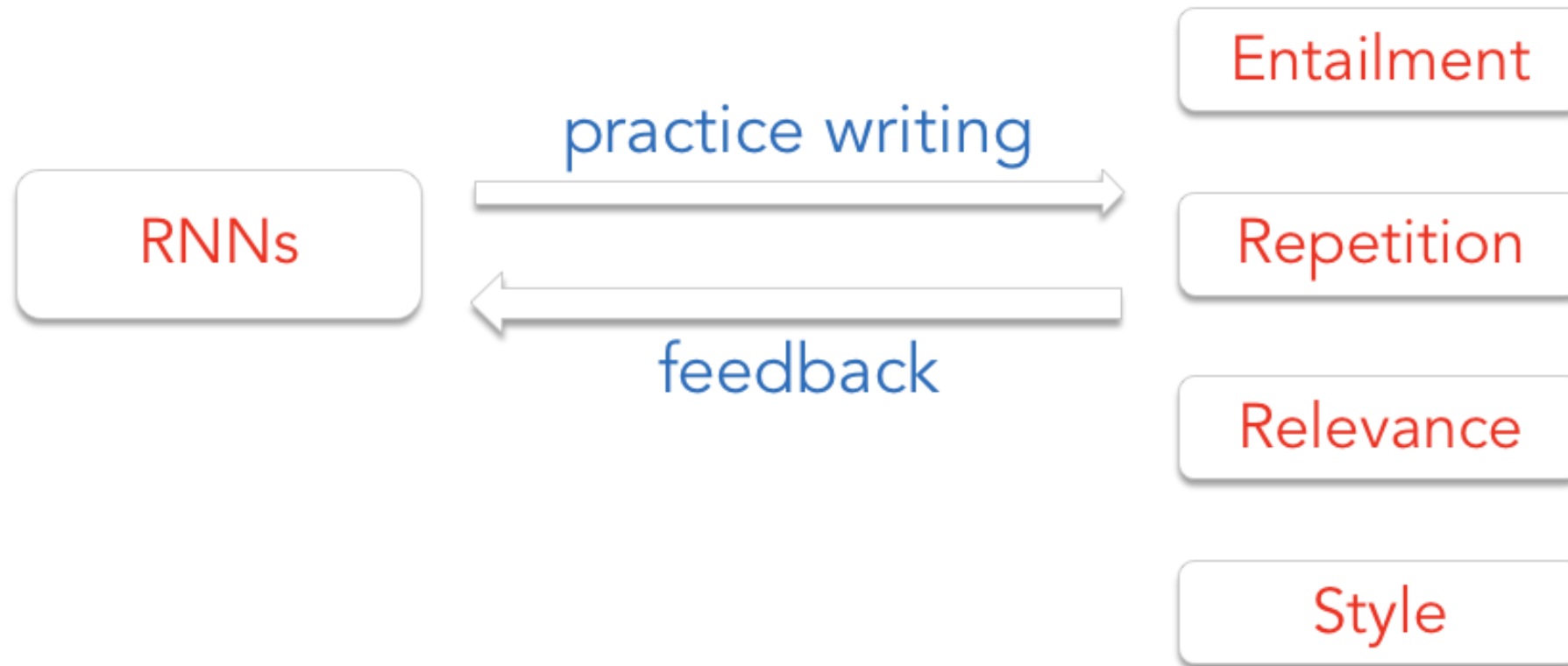
Be Informative

- Train discriminator to distinguish between references and LM generated text, but using a model to detect repetition patterns

$$\sum_{\substack{(\mathbf{x}, \mathbf{y}_{\text{human}}) \in D, \\ \mathbf{y}_{\text{distractor}} \sim P_{\text{LM}}|\mathbf{x}}} \log \sigma(s_{\text{rep}}(d(\mathbf{x}||\mathbf{y}_{\text{human}})) - s_{\text{rep}}(d(\mathbf{x}||\mathbf{y}_{\text{distractor}})))$$

Discriminators inspired by Grice's Maxims

Truthful, Informative, Relevant, Clear



Relevance Module

Be Relevant

- Compare true continuation to one randomly selected from the data
- Train discriminator to rank correct continuation above incorrect one
- Convolutional neural network encodes the initial text \mathbf{x} and candidate continuation \mathbf{y}

$$\sum_{\substack{(\mathbf{x}, \mathbf{y}_{\text{human}}) \in D \\ \mathbf{y}_{\text{distractor}} \sim D_{\mathbf{y}}}} \log \sigma(s_{\text{rel}}(\mathbf{x}, \mathbf{y}_{\text{human}}) - s_{\text{rel}}(\mathbf{x}, \mathbf{y}_{\text{distractor}}))$$

Relevance Module

Given:

When the bodies were loaded and the cortege assembled outside the gates, the Larisien leaders said their farewells.

The LM continues:

"I'm sorry," she said. "I'm so sorry," he said.

Our model continues:

They didn't have to wait long. It was time for them to leave.

Style Module

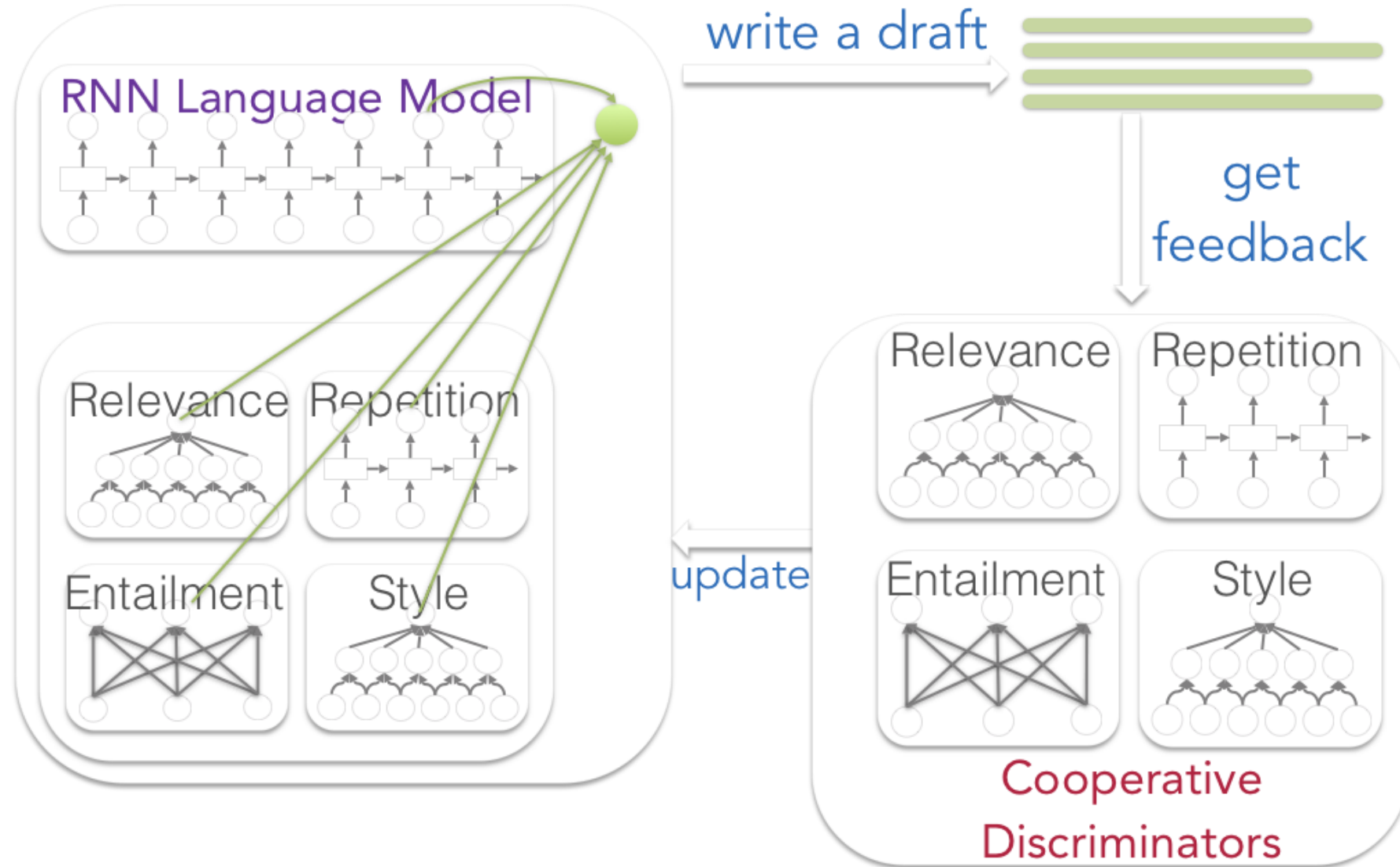
Be Clear

- Compare the reference generation to a generation sampled from the language model, ignoring the initial text

$$\sum_{\substack{(\mathbf{x}, \mathbf{y}_{\text{human}}) \in D, \\ \mathbf{y}_{\text{distractor}} \sim P_{\text{LM}}|\mathbf{x}}} \log \sigma(s_{\text{sty}}(\mathbf{y}_{\text{human}}) - s_{\text{sty}}(\mathbf{y}_{\text{distractor}}))$$

Learning to Write

Holtzman, Buys et al., ACL 2018



Learning to Write

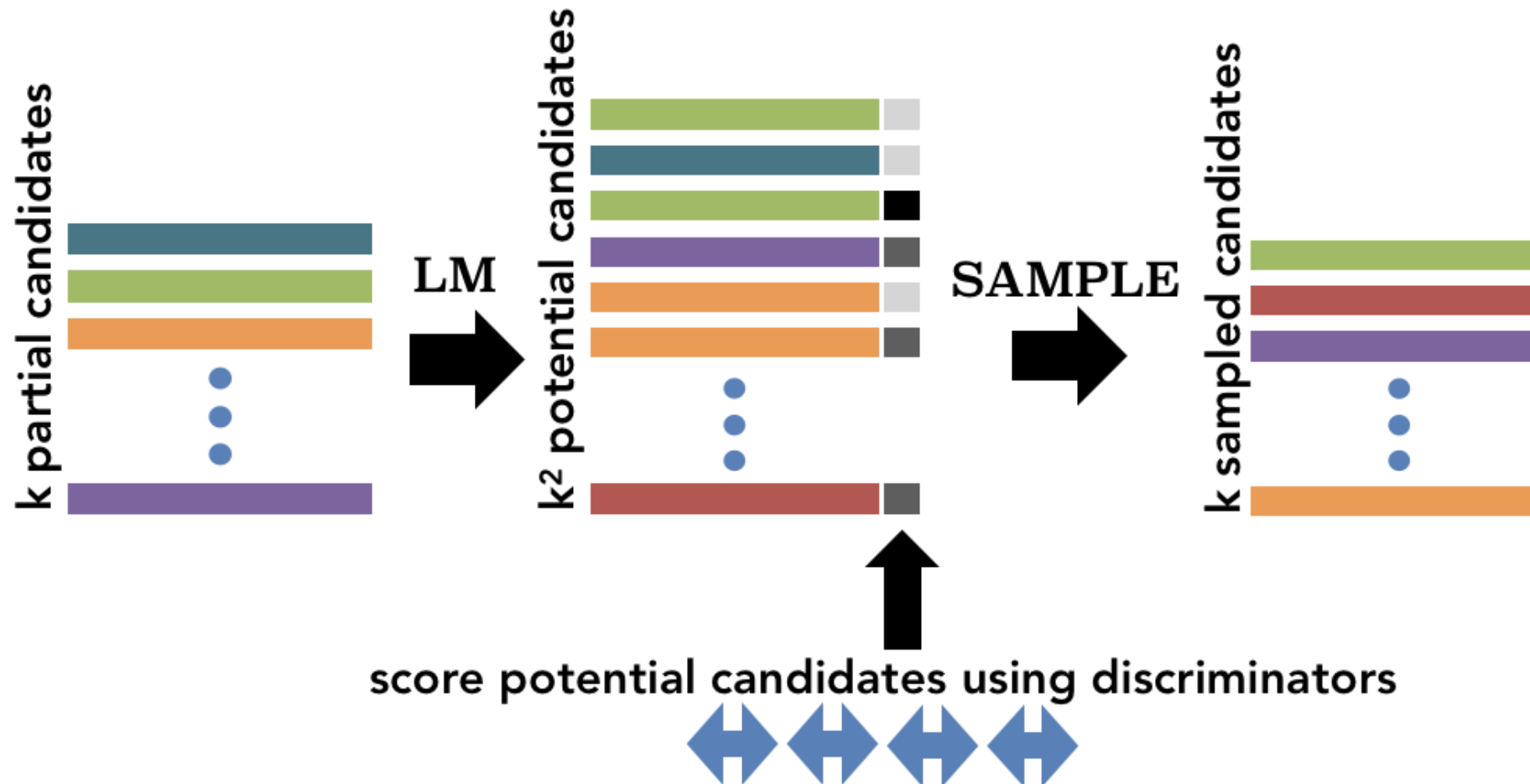
- Decoding objective is a weighted combination of base LM score and discriminator scores.

$$f_{\lambda}(\mathbf{x}, \mathbf{y}) = \log(P_{\text{LM}}(\mathbf{y}|\mathbf{x})) + \sum_k \lambda_k s_k((x), \mathbf{y})$$

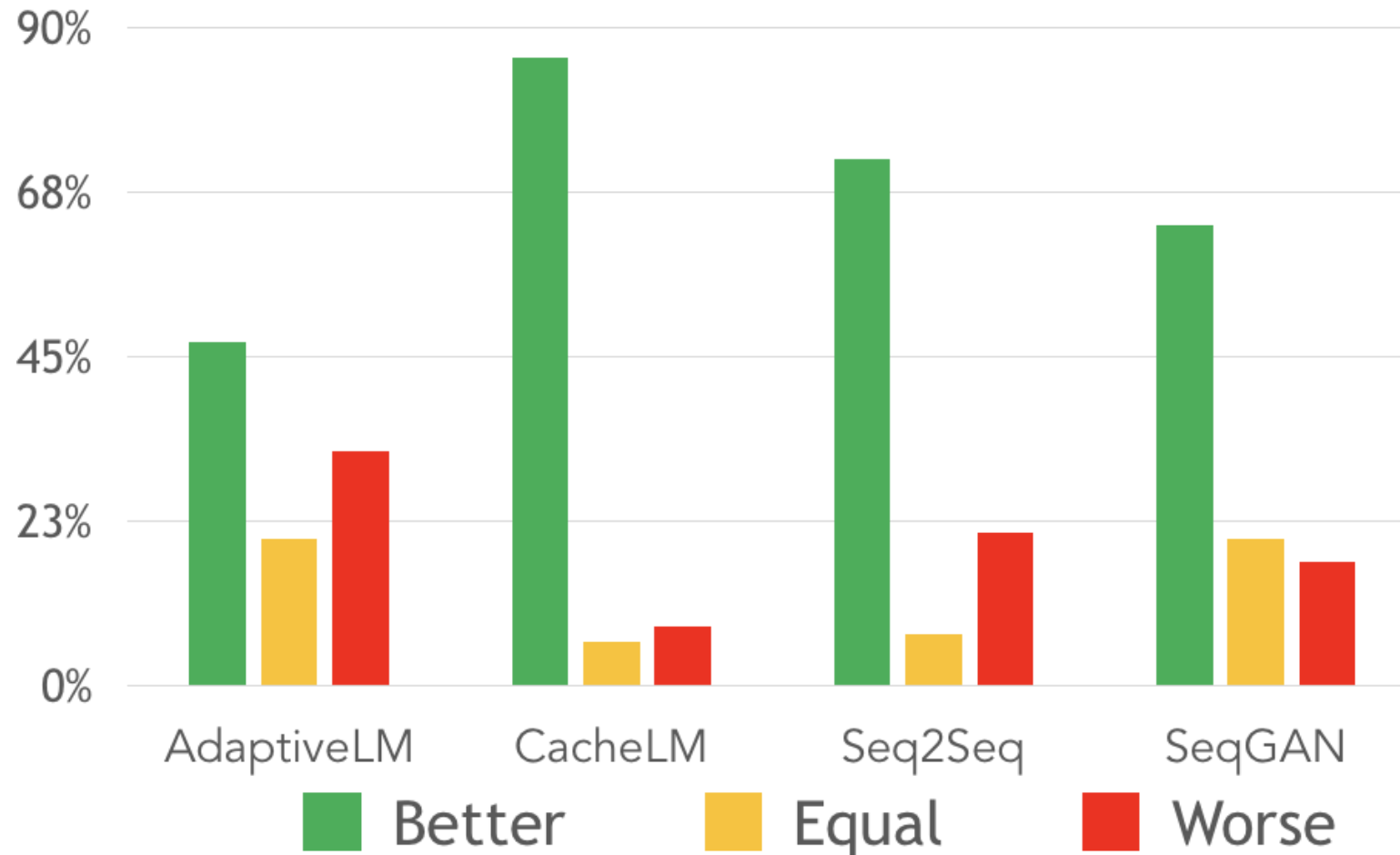
- Learn the mixture coefficients that lead to the best generations

Learning to Write: Generation

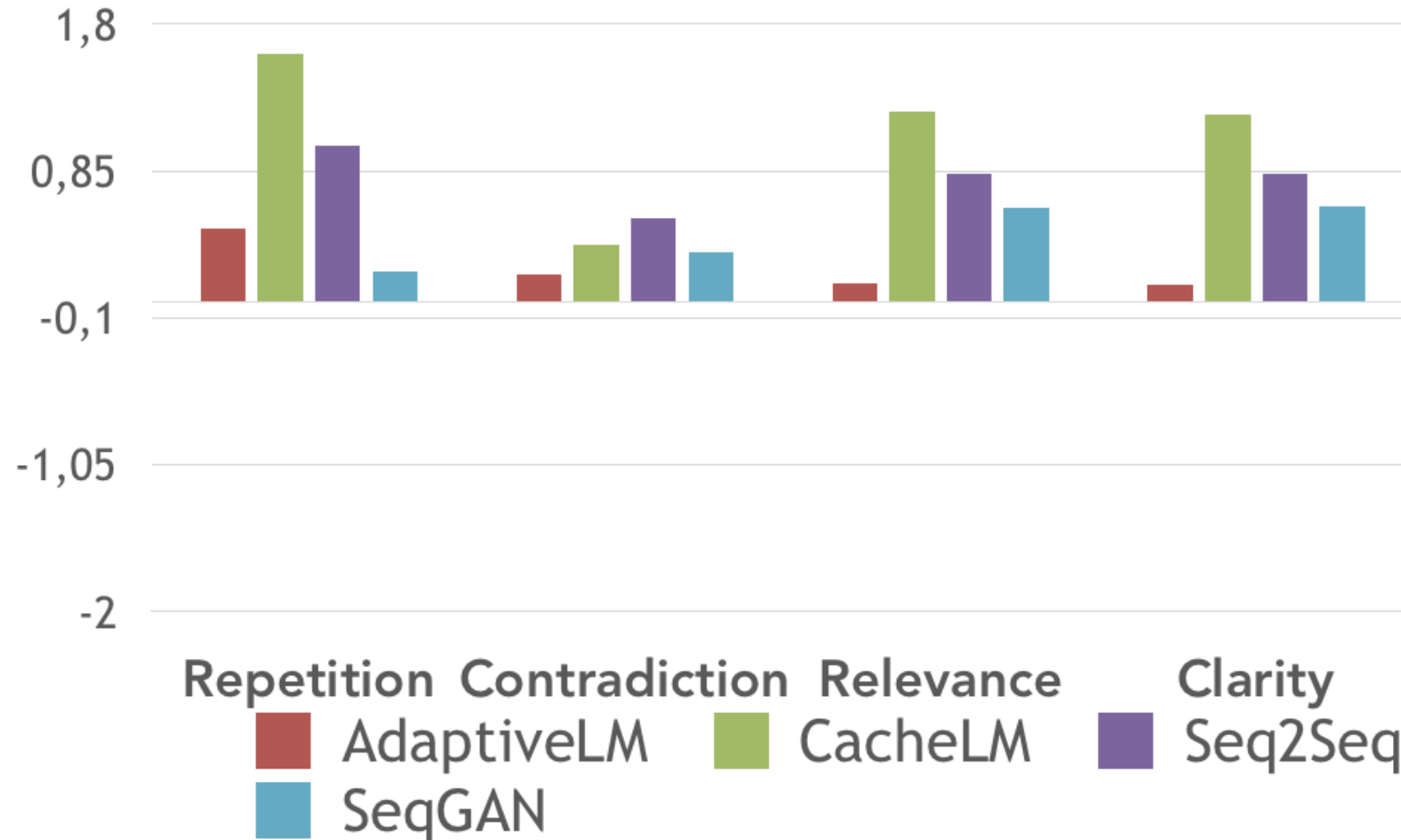
- Top-k sampling with (stochastic) beam search over decoding objective



Human Evaluation: L2W vs Baselines



Human Evaluation: L2W vs Baselines



Example Generations

It was as though the entire room had been thrown back at him, like a rag doll. His face was contorted in pain, his eyes wide and terrified. He stood there for a moment looking up at the ceiling and then at the floor. "don't worry about it," he whispered. "I 'll take you."

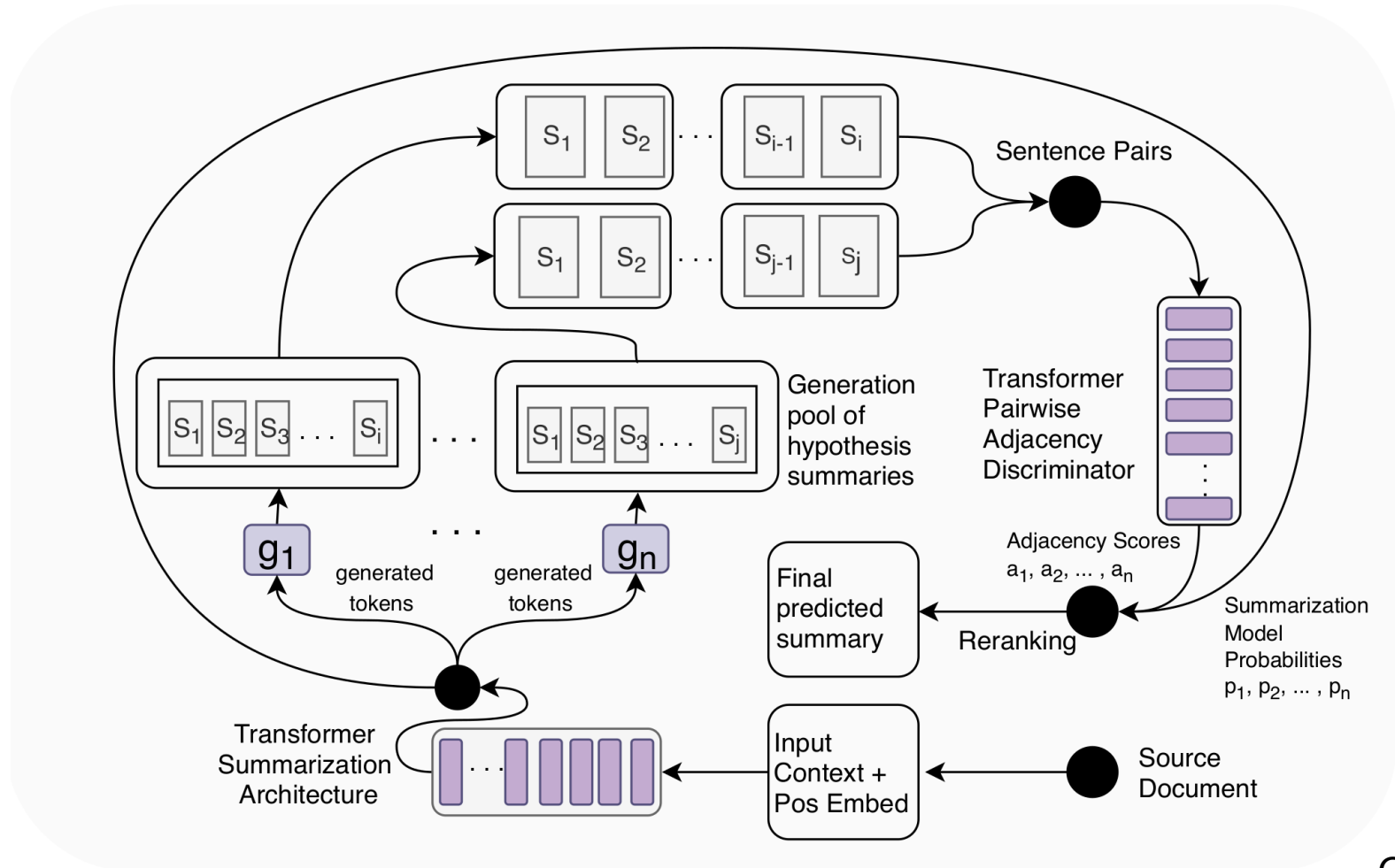
Example Generations

It had been a long day for him, and he was grateful for her presence at the funeral home. She had gone back to her own room. There were a few nurses and a nurse who had come to the hospital. She had been so patient with him that she had not been able to tell him what had happened. The doctor did.

Example Generations

She didn't know what else to do. She looked up at her father and saw him staring at her. He was looking down at her, his blue eyes glittering in the dim light as he looked at her. "are you okay?" He asked.

Cooperative Generator–Discriminators for Narrative Summarization

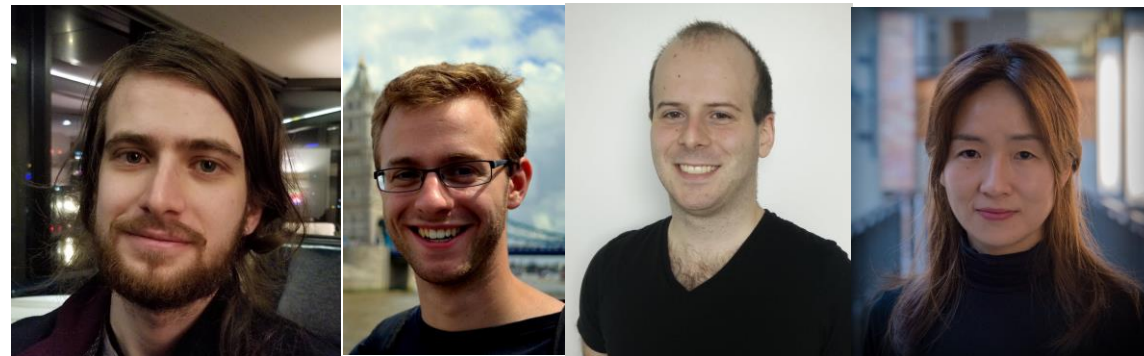


Conclusion

- Learning to Write produces more **diverse** and **coherent** generations
- The framework **prevents** the model from **ignoring the context** and from the generation **quality degrading**
- Better modelling is complementary to better decoding

Bibliography

- The Curious Case of Neural Text Degeneration. *Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes and Yejin Choi*. **ICLR 2020**.
- Learning to Write with Cooperative Discriminators. *Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub and Yejin Choi*. **ACL 2018**.



Thank you

jbuys@cs.uct.ac.za

www.janmbuys.com