



Why NLU doesn't generalize to NLG

Yejin Choi

Paul G. Allen School of Computer Science & Engineering

W UNIVERSITY *of* WASHINGTON

& Allen Institute for Artificial Intelligence

"In its current form..."

"neural"

Why NLU doesn't generalize to NLG

"well"

NLG depends less on NLU

- Pre-DL, NLG models often started with NLU output.
- Post-DL, NLG seems less dependent on NLU.
 - What brought significant improvements in NLG recent years isn't so much due to better NLU (tagging, parsing, co-ref'ing, QA'ing).
- In part because end-to-end models work better than pipeline models.
 - It's just seq-2-seq with attention!

NLG depends heavily on Neural-LMs

- Conditional models:
 - Sequence-to-sequence models

$$p(x_1, \dots, x_n | context) = \prod_i p(x_i | x_1, \dots, x_{i-1}, context)$$

- Generative models:
 - Language models

$$p(x_1, \dots, x_n) = \prod_i p(x_i | x_1, \dots, x_{i-1})$$

Works amazingly well
for MT, speech reg,
image captioning, ...

Neural generation was not part of the winning recipe for the Alexa challenge 2017.

however,
neural generation can be brittle

"even templated baselines exceed the performance of these neural models on some metrics ..."

- Wiseman et al., EMNLP 2017

neural generation can be brittle
(no adversary necessary)

All in all, I would highly recommend this hotel to
anyone who wants to be in the heart of the action

neural generation can be brittle
(no adversary necessary)

All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action, and want to be in the heart of the action. If you want to be in the heart of the action, this is not the place for you. However, If you want to be in the middle of the action, this is the place to be.

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

neural generation can be brittle (no adversary necessary)

All in all, I would highly recommend this hotel to anyone who wants **to be in the heart of the action**, and want **to be in the heart of the action**. If you want **to be in the heart of the action**, this is not the place for you. However, If you want to be in the middle of the action, this is the place to be.

repetitions...

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

neural generation can be brittle (no adversary necessary)

All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action, and want to be in the heart of the action. If you want to be in **the heart of the action, this *is not the place*** for you. However, If you want to be in **the middle of the action, this *is the place to be***.

contradictions...

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

neural generation can be brittle (no adversary necessary)

All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action,
and want to be in the heart of the action. If you want
to be in the heart of the action, this is not the place
for you. However, If you want to be in the middle of
the action, this is the place to be.

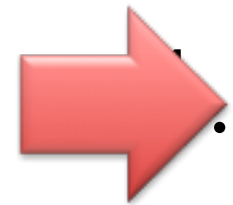
generic, bland, lack of details

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

natural language in,
unnatural language out.
why?

- Not enough depth?
- Not enough data?
- Not enough GPUs?
- Even with more depth, data, GPUs, I'll speculate that current LM variants are not sufficient for robust NLG

Two Limitations of LMs



1. Language models are *passive* learners

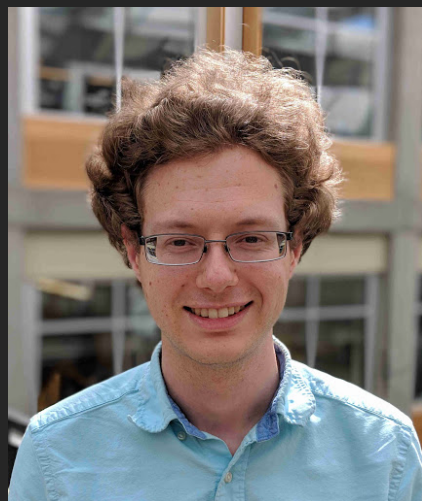
- one can't learn to write just by reading
- even RNNs need to "*practice*" writing

2. Language models are *surface* learners

- we also need **world** models
- the **latent process** behind language

Learning to Write with Cooperative Discriminators

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, Yejin Choi @ ACL 2018



neural generation can be brittle
(no adversary necessary)

All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action, and want to be in the heart of the action. If you want to be in the heart of the action, this is not the place for you. However, If you want to be in the middle of the action, this is the place to be.



Symptoms?

- Often goes into a repetition loop.
- Often contradicts itself.
- Generic, bland, and content-less.

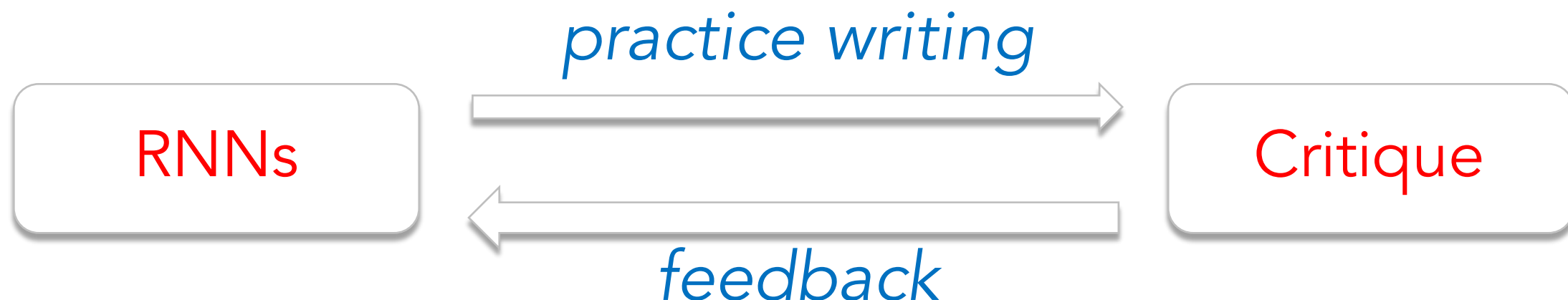
Causes?

- Learning objective isn't quite right
 - people don't write to maximize the probability of the next token
- Long context gets ignored
 - “explained away” by more appealing short-term context (Yu et al., 2017)
- Inductive bias isn't strong enough
 - LSTMs/GRUs architectures not sufficient for learning discourse structure

Solution:

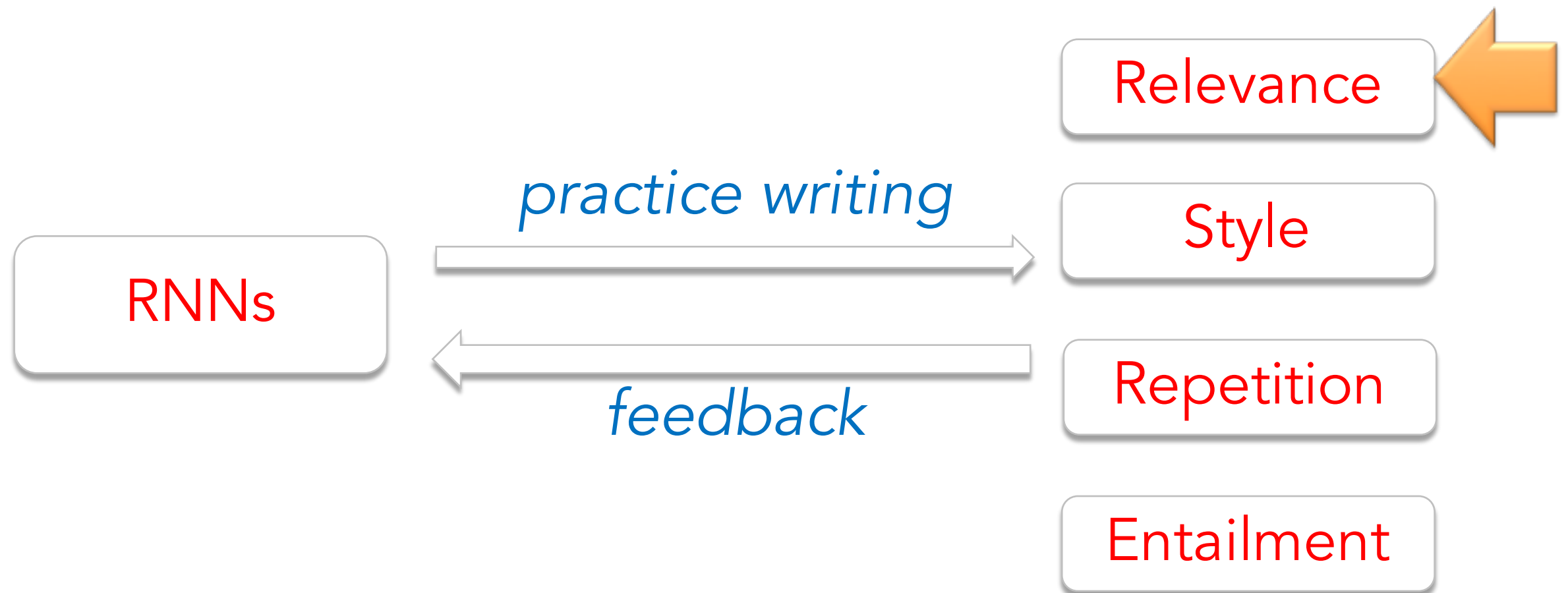
“Learning to Write by Practice”

- let RNNs practice writing
- A committee of critiques compare RNN text to human text
- RNNs learn to write better with the guidance from the cooperative critiques



Discriminators inspired by Grice's Maxims

Quantity, Quality, Relation, Manner



Relevance Module

Given:

We had an inner room and it was quiet.

The base LM
continues...

**The staff was very
friendly, helpful, and
polite.**

L2W continues...

**There was a little noise
from the street, but
nothing that bothered
us.**

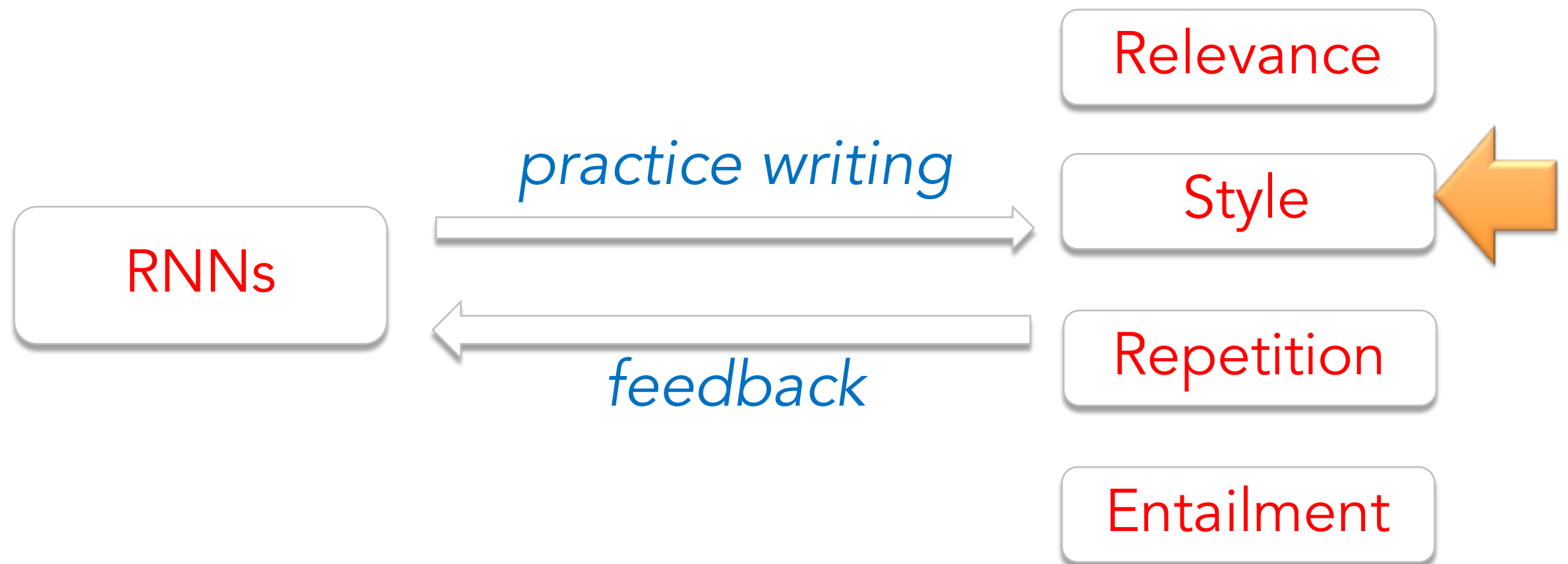
Relevance Module

- Both continuations are fluent, but the true continuation will be more relevant.
- A convolutional neural network encodes the initial text \mathbf{x} and candidate continuation \mathbf{y} .
- Trained to optimize a ranking loss:

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

Discriminators inspired by Grice's Maxims

Quantity, Quality, Relation, Manner



Style Module

LM

"It's time to go,"
the woman said.
"It 's time to go."
She turned back
to the others. "I'll
be back in a
moment." She
nodded.

L2W

They didn't speak at all.
Instead they stood staring at
each other in the middle of the
night. It was like watching a
movie. It felt like an eternity
since the sky above them had
been lit up like a Christmas
tree. The air around them
seemed to move and breathe.

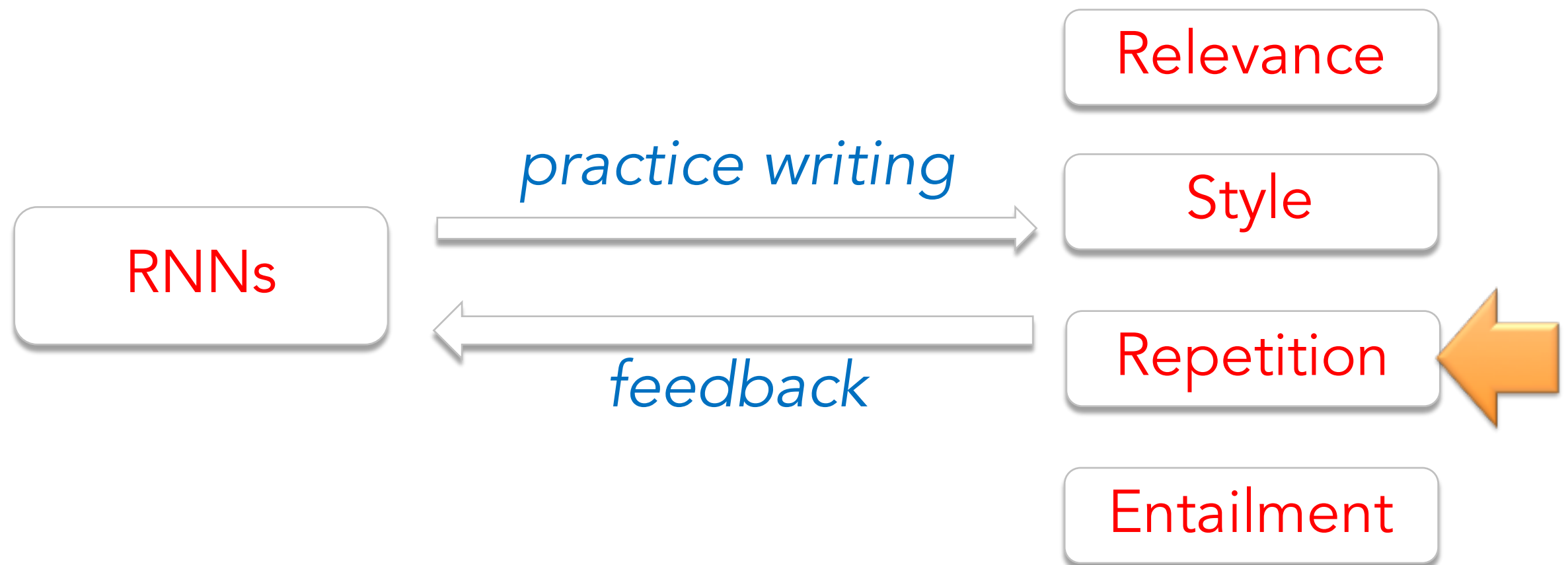
Style Module

Convolutional architecture and loss function similar to the relevance module, but conditions only on the generation, not on the initial text.

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

Discriminators inspired by Grice's Maxims

Quantity, Quality, Relation, Manner



Repetition Module

LM:

He was dressed in
a white t-shirt,
blue jeans, and a
black t-shirt.

L2W:

His eyes were a shade
darker and the hair on
the back of his neck
stood up, making him
look like a ghost.

Repetition Module

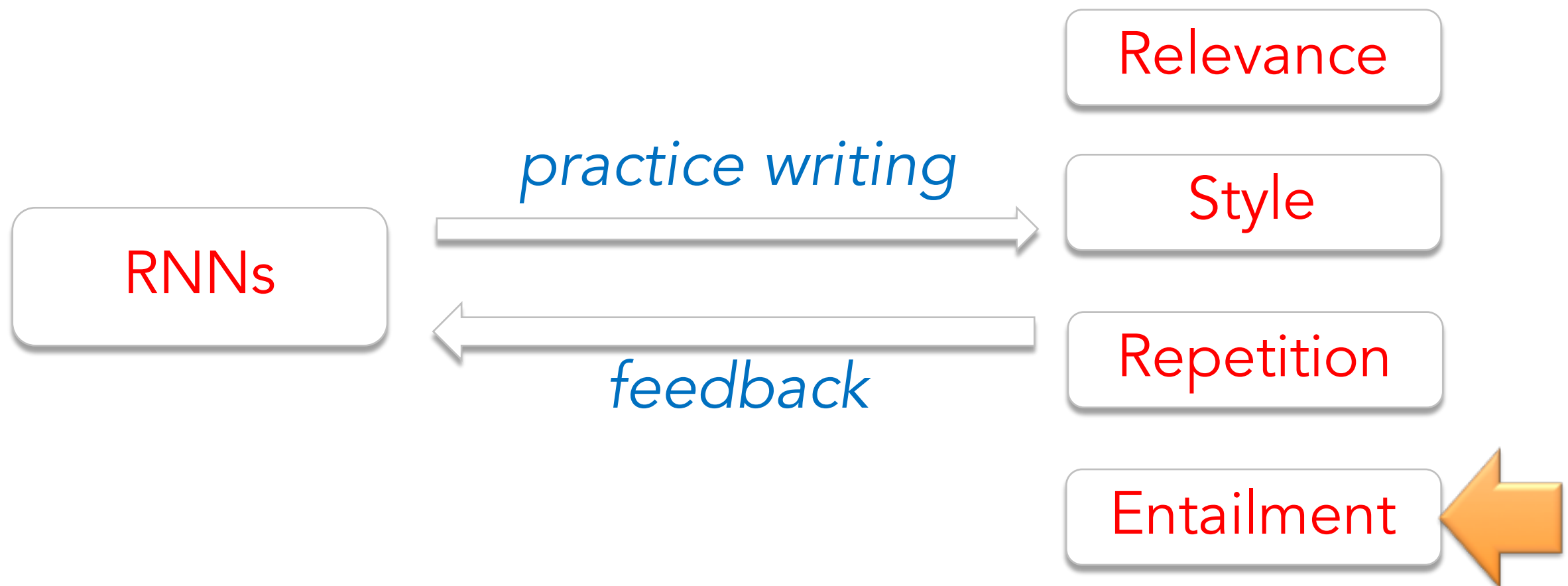
- Train an RNN-based discriminator to distinguish between LM generated text and references, conditioned only on these similarity sequences:

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

Parameterizing undesirable repetition through embedding similarity, instead of placing a hard constraint of not repeating ngrams (Paulus et al., 2018)

Discriminators inspired by Grice's Maxims

Quantity, Quality, Relation, Manner



Entailment Module

I loved the in-hotel restaurant!



There was an in-hotel restaurant.

Entailment Module

I loved the in-hotel restaurant!



CONTRADICT

The closest restaurant was ten miles away.

In summarization, it's "entailment" that we want to encourage between input and output
- Pasunuru and Bansal, NAACL 2018

I loved the in-hotel restaurant!



NEUTRAL

It's a bit expensive, but well worth the price!

Entailment Module

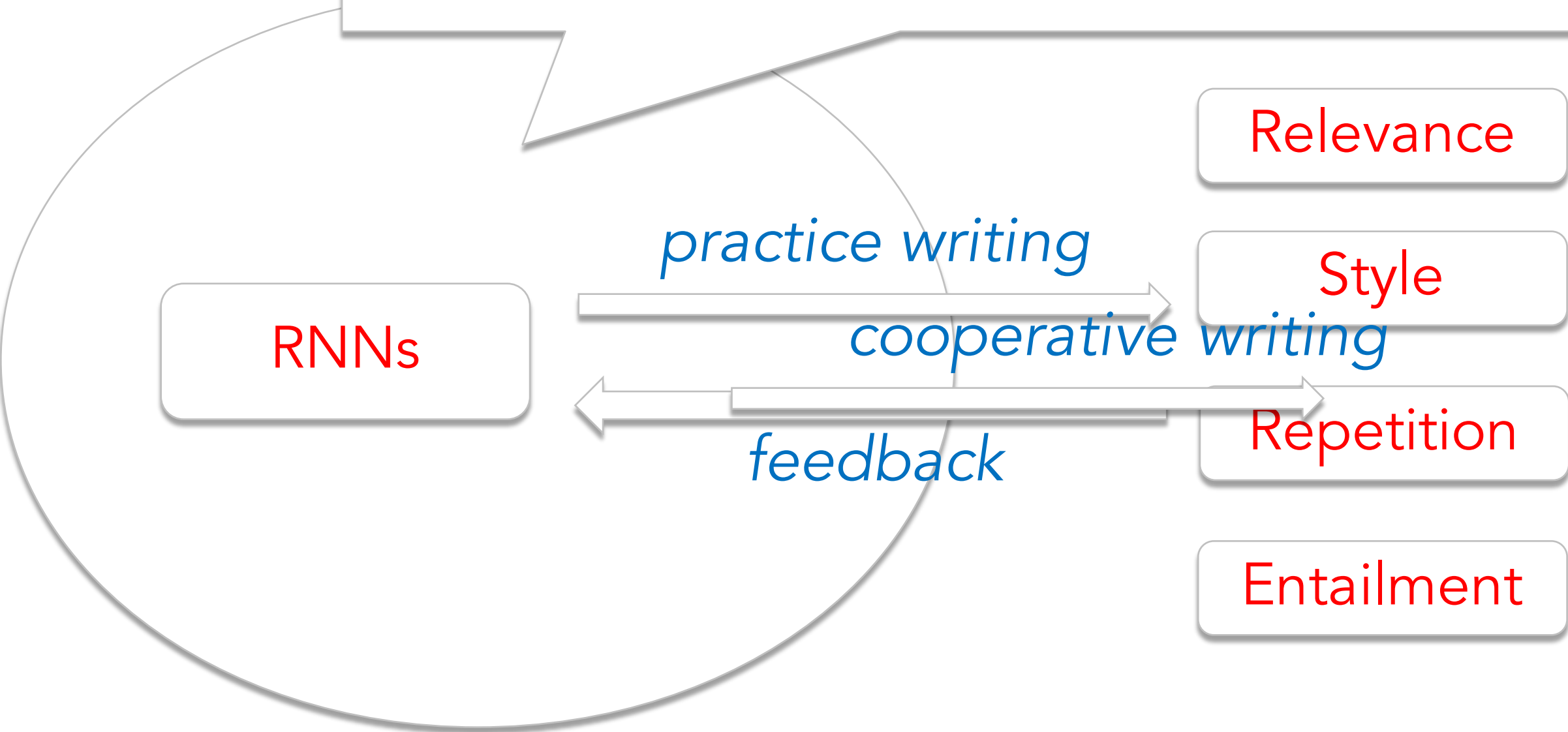
- Compare candidate sentence to each previous sentence, and use minimum probability of the *neutral* category—neither entailing nor contradiction.
- Trained on SNLI + MNLI dataset (Bowman et al., 2015, Williams et al., 2017) using the decomposable attention model (Parikh et al., 2016)

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

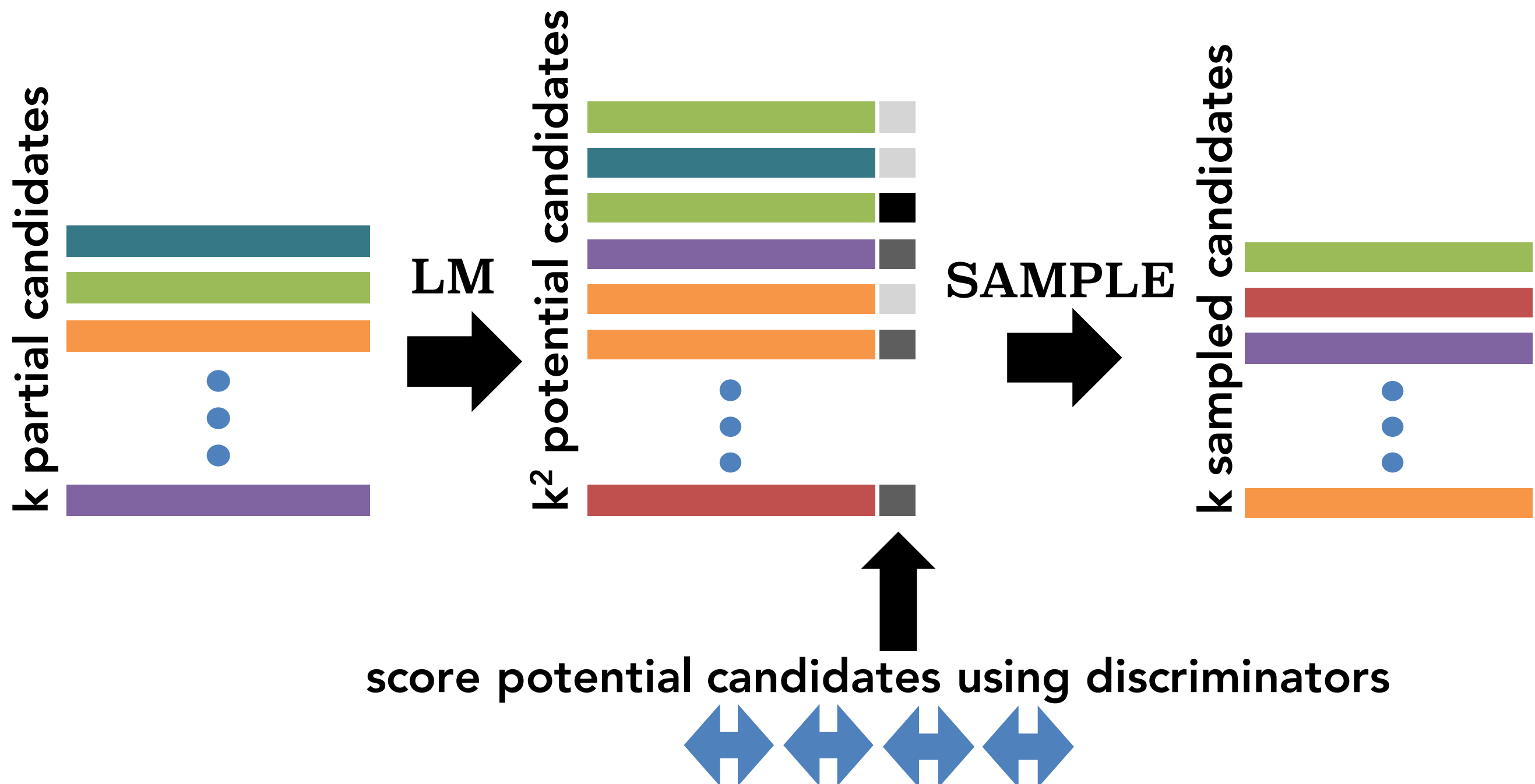
where $S(\mathbf{x})$ are the initial sentences and $S(\mathbf{y})$ are the completed sentences.

Integration of NLG with NLU!

- *NLU of unnatural (machine) language*
- *NLU without formal linguistic annotations*



Generation with Cooperative Discriminators



Learning to Write with Cooperative Discriminators

- The decoding objective function is a weighted combination of the base LM score and discriminator scores.
 - “Product of experts” (Hinton 2002)

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

- We **learn** the mixture coefficients that will lead to the best generations.
- Loss: $(f_{\lambda}(\text{human}) - f_{\lambda}(\text{machine}))^2$

Datasets

- TorontoBook Corpus
 - 980 million words, amateur fiction.
- TripAdvisor
 - 330 million words, hotel reviews.

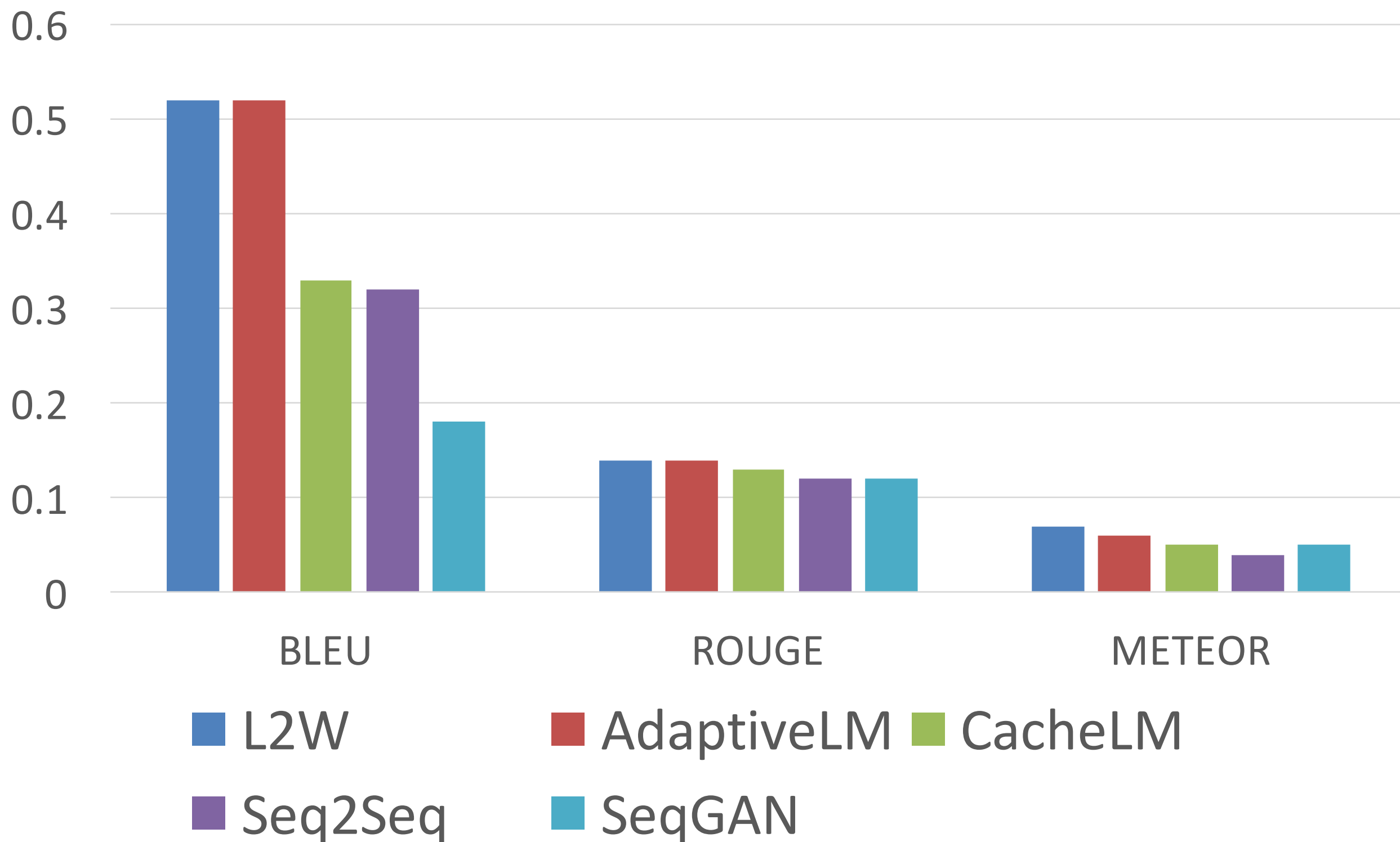
Input & output setup:

- use 5 sentences as context,
- generate the next 5 sentences.

Baselines

- AdaptiveLM
- CacheLM
- Seq2Seq
- SeqGAN

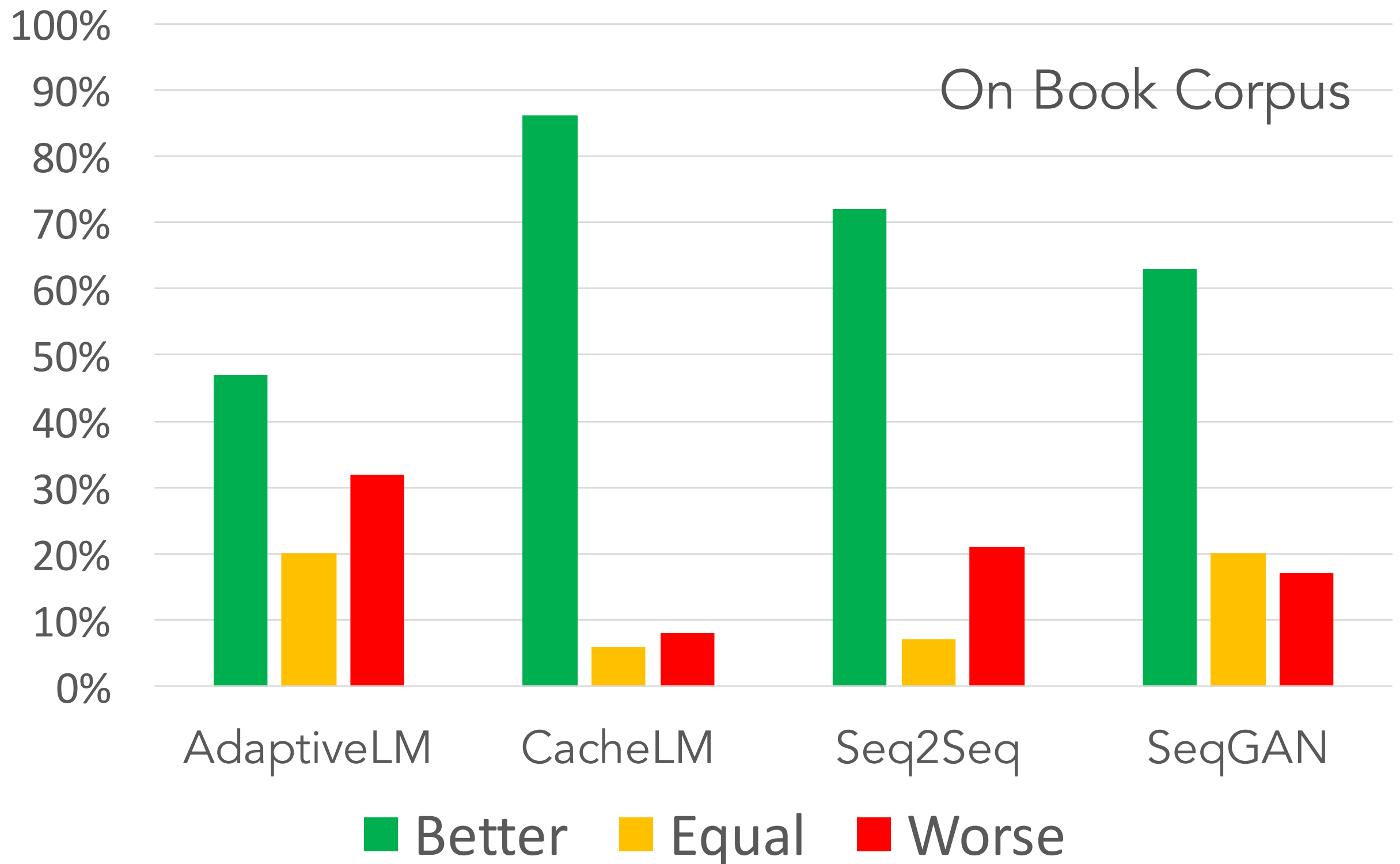
ngram based evaluation



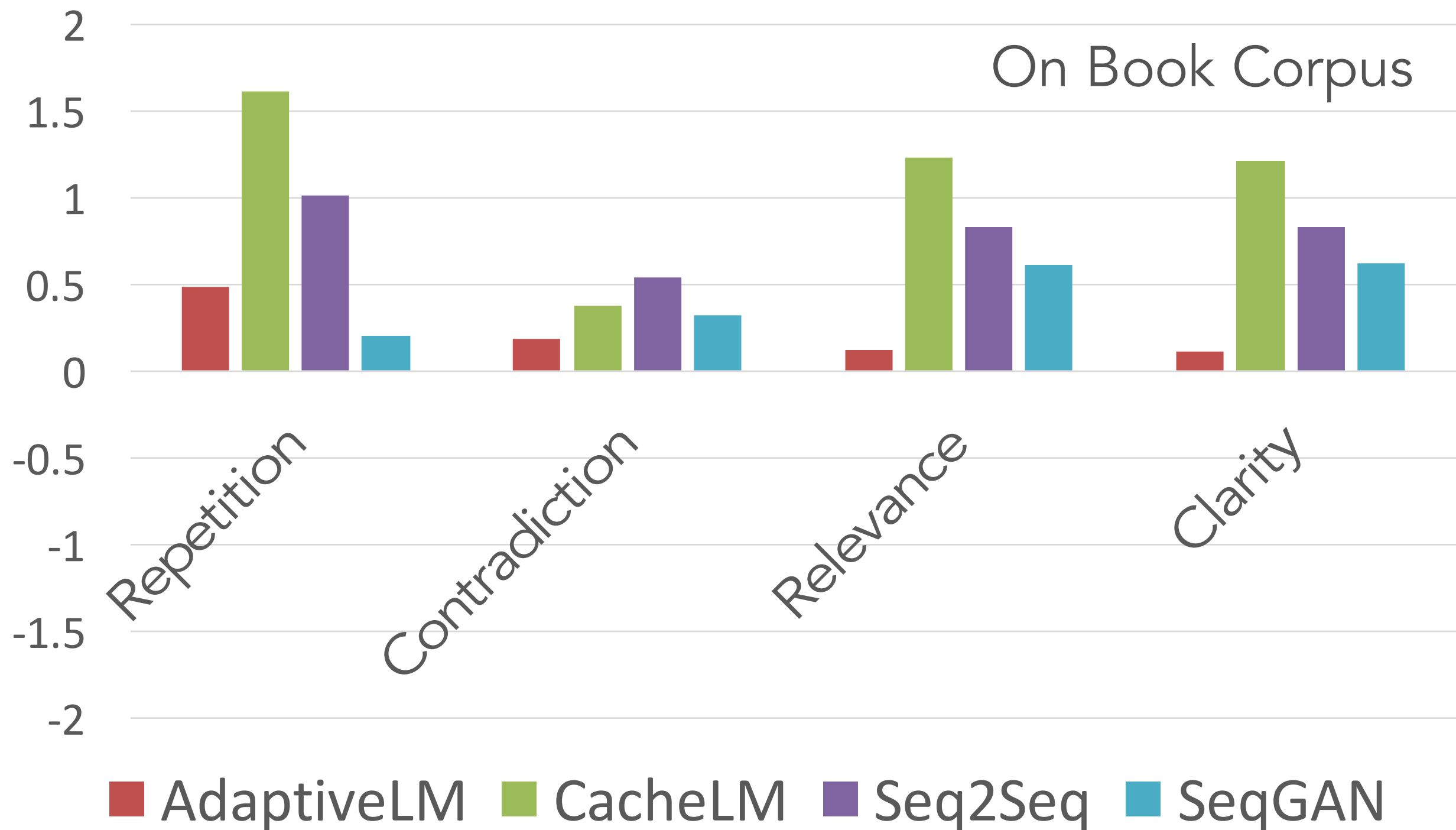
N-gram overlap only measures
the surface pattern matching

Not the true quality of the
generated text

Human Eval: L2W vs. *X*



Human Eval: L2W vs. X



L2W vs. Seq2Seq

Seq2Seq:

She looks up at the sky. The light shines brighter. The light flickers. The light flickers. The light flickers.

L2W:

She's staring at the floor. She's shaking her head in wonder. "What do you mean?" "Don't you know what it's like in your dreams?" Nora asks.

L2W vs. SeqGAN

SeqGAN:

We're going to leave.
When he ended up
here, he was here. Not
not really. It was... It
was a hard question.

L2W:

There was only one way to
find out. The door swung
open. Gregor stood in the
middle of the room,
staring up at the ceiling.
His eyes were wide and
his breathing was shallow.
"What's happening?"


Human or Machine?

Our maid service was prompt and efficient (the beds weren't made-up until late afternoon—but we had plenty of towels, etc!). Also, there's a free shuttle to/from Walt Disney World, Universal or Orlando premium outlets, plus there's a 24-hour Walgreens nearby to stock up on snacks for the room, as well as an internet cafe that's open 24 hours a day, which made it convenient for us. I'd recommend this property to families and business travelers who don't mind spending a bit more money, but would probably stay somewhere else next time. It's very nice with friendly staff, good pool facilities, and excellent on-site dining options. However, the prices at the parks are ridiculously inflated.

Two Fundamental Issues with LMs

1. Language models are *passive learners*

- one can't learn to write just by reading
- even RNNs need to "*practice*" writing

 2. Language models are *surface learners*

- we also need **world** models
- the **latent process** behind language

significant performance gaps across different tasks

Why no one reports
super-human performance on

making a conversation,
summarizing a document,
composing/replying to emails,
identifying fake news

?

Not robust if given

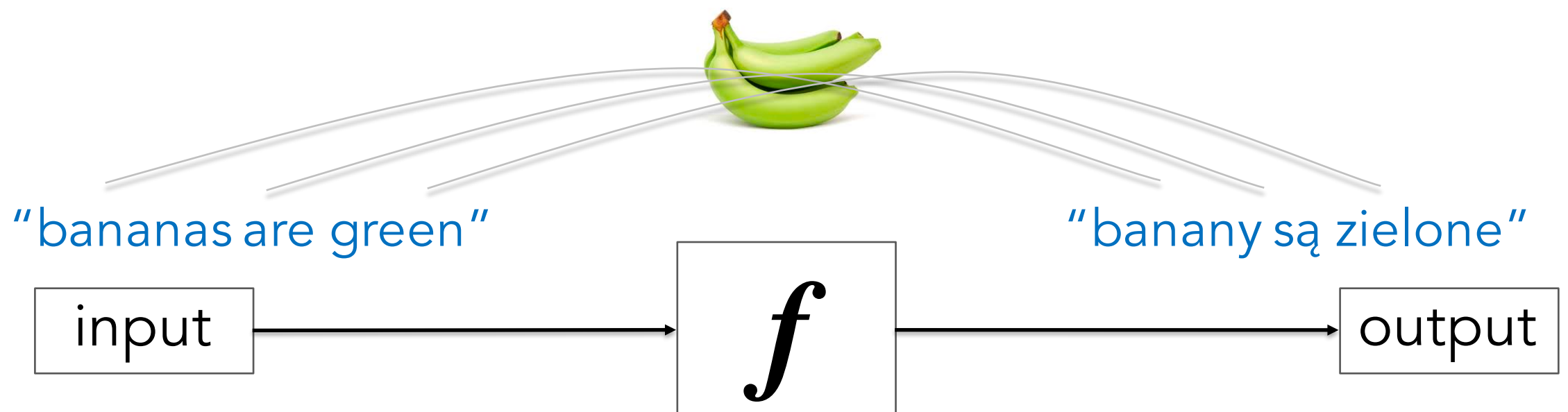
unfamiliar,
out-of-domain
or adversarial
examples

(Jia et al., 2017, Belinkov et al.,
2018)

super-human performance on object recognition

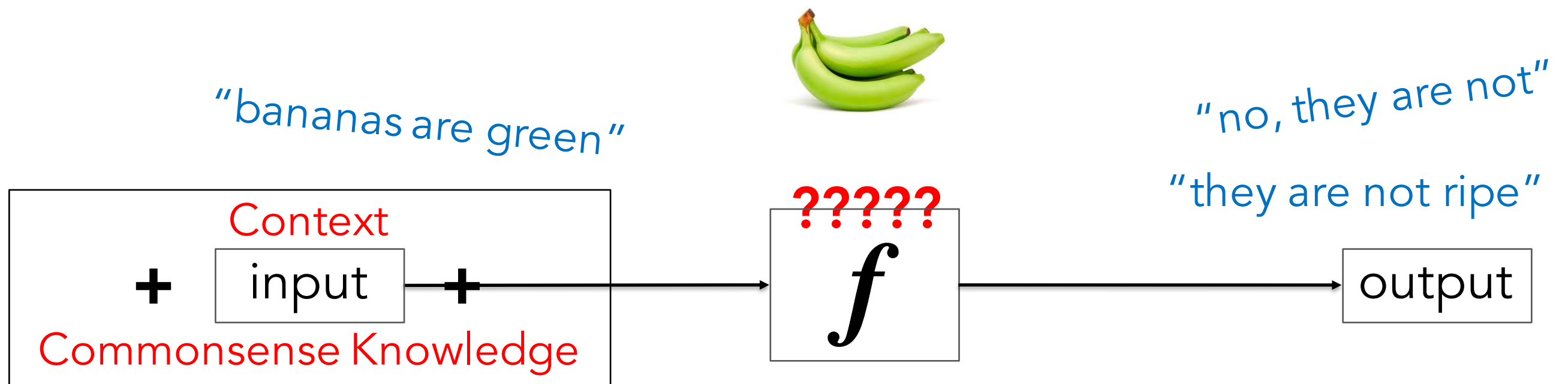
Why significant performance gaps

- Type 1 (shallow NLU):
 - Strong alignment between input and output
 - Surface pattern matching
- Type 2 (deep NLU):

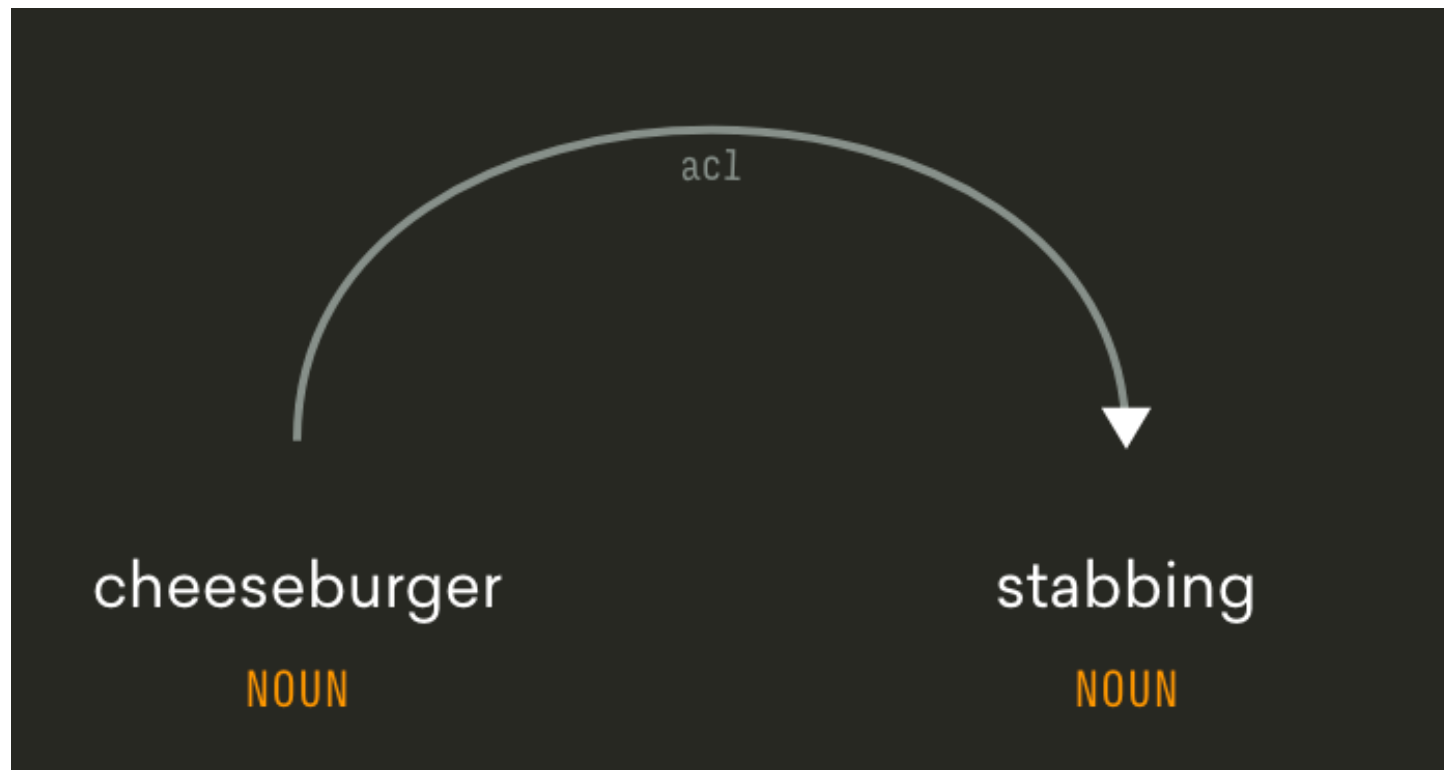


Why significant performance gaps

- Type 1 (shallow NLU):
 - Strong alignment between input and output
 - Surface pattern matching
- Type 2 (deep NLU):
 - Weak alignment between input and output
 - Abstraction, cognition, reasoning
 - Requires knowledge, especially commonsense knowledge



Reading between the Lines

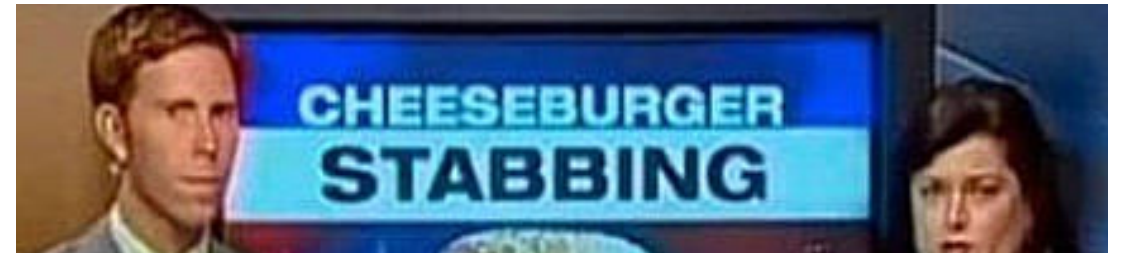


"CHEESEBURGER STABBING"

- Someone stabbed a cheeseburger?
- A cheeseburger stabbed someone?
- A cheeseburger stabbed another cheeseburger?
- Someone stabbed someone else over a cheeseburger?

Reading between the Lines

➔ Reading between the lines



- **Physical Commonsense:** not possible to stab using a burger
- **Social Commonsense:** stabbing someone is bad

what is not said



"CHEESEBURGER STABBING"

- Someone stabbed a cheeseburger?
- A cheeseburger stabbed someone?
- A cheeseburger stabbed another cheeseburger?
- Someone stabbed someone else over a cheeseburger?

Types of Knowledge

Information Extraction

Encyclopedic knowledge

- Who is the president of which country and born in what year...

Naïve Physics

Commonsense knowledge

- It's not possible to stab someone using a cheeseburger
- Stabbing a cheeseburger is not newsworthy...
- Stabbing someone is generally immoral

Social Norms

Commonsense

- Searching “commonsense” from ACL anthology
 - Most papers are either from 80s or from the past few years

Position Paper on Common-sense and Formal Semantics

Geoffrey Nunberg
Xerox PARC and CSLI, Stanford

1. A philological excursus

I'm not sure what I'm doing on this panel, but I thought it would be helpful if we could start at the beginning. It's interesting to note that both the dictionary and common sense were eighteenth-century inventions. This is no coincidence; in fact, it's entirely appropriate that the most celebrated

Recent (Commonsense) Challenges

- Winograd Schema Challenge (Levesque et al., 2014)

The trophy would not fit in the brown suitcase because it was too big. What was too big?

Answer 0: the trophy

Answer 1: the suitcase

- Commonsense Story Cloze (Mostafazadeh et al., 2016)
- Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011)
- LAMBADA Story Understanding Dataset (Parperno et al., 2016)

➔ Models based on surface pattern matching fail on these tasks

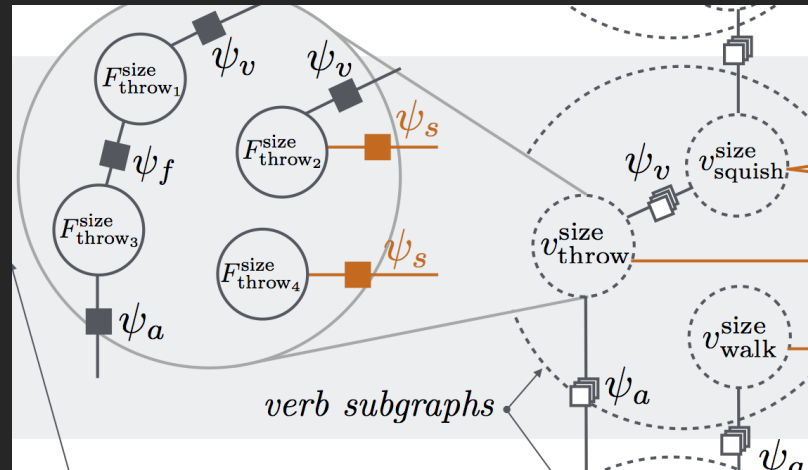
➔ Brute force large-scale training does not seem promising

Revisiting Commonsense

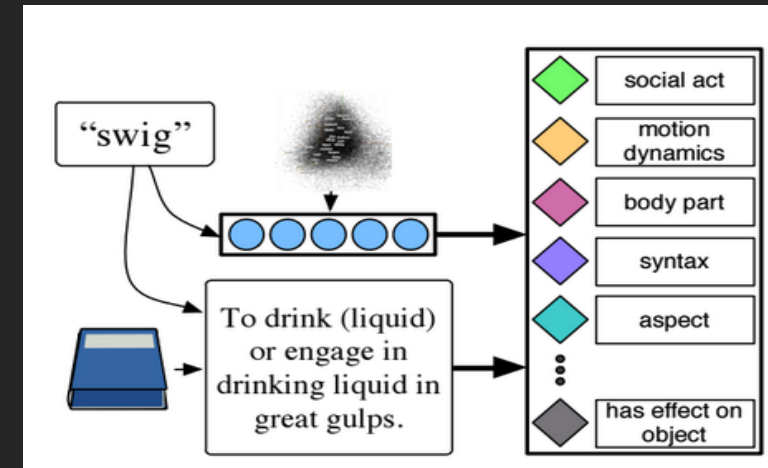
I was told not to use the word “commonsense”...

Past failures (in 70s – 80s) are inconclusive

- weak computing power
- not much data
- no crowdsourcing
- not as strong computational models
- not ideal conceptualization / representations



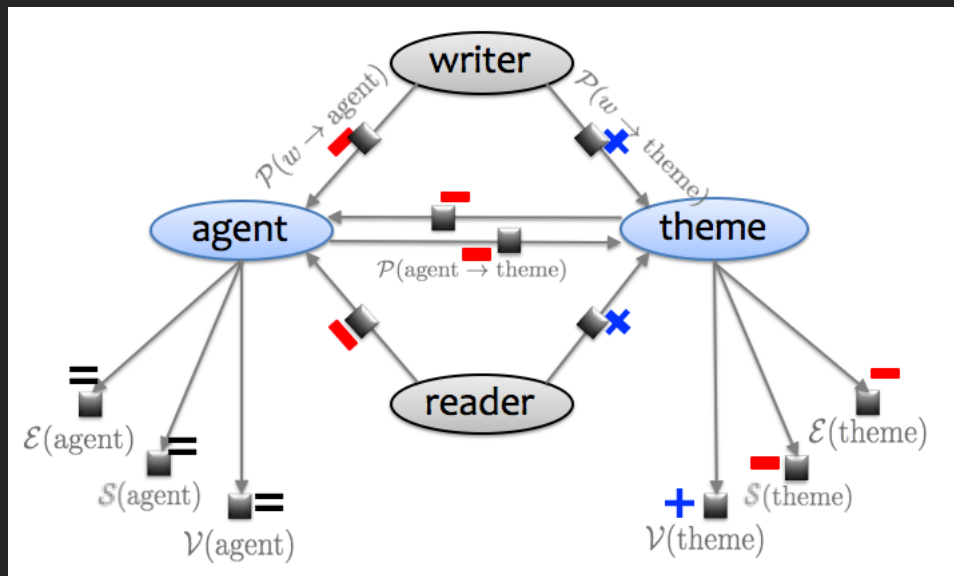
VerbPhysics (ACL 2017)



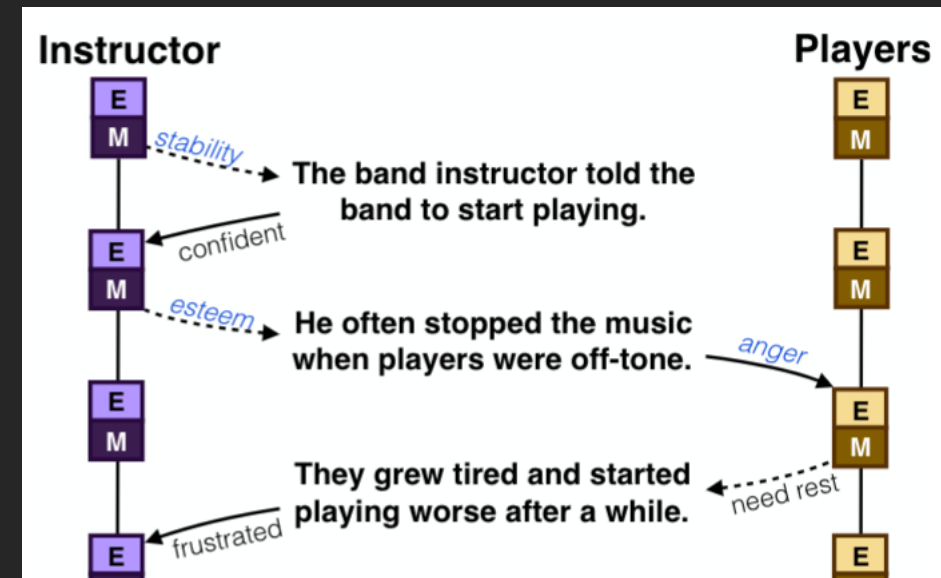
Zero-shot activity recognition with verb attribute induction (EMNLP 2017)

Physical commonsense

- Zero-shot / few-shot learning
- Language and vision
- Language and robotics



Connotation Frames (ACL 2016)

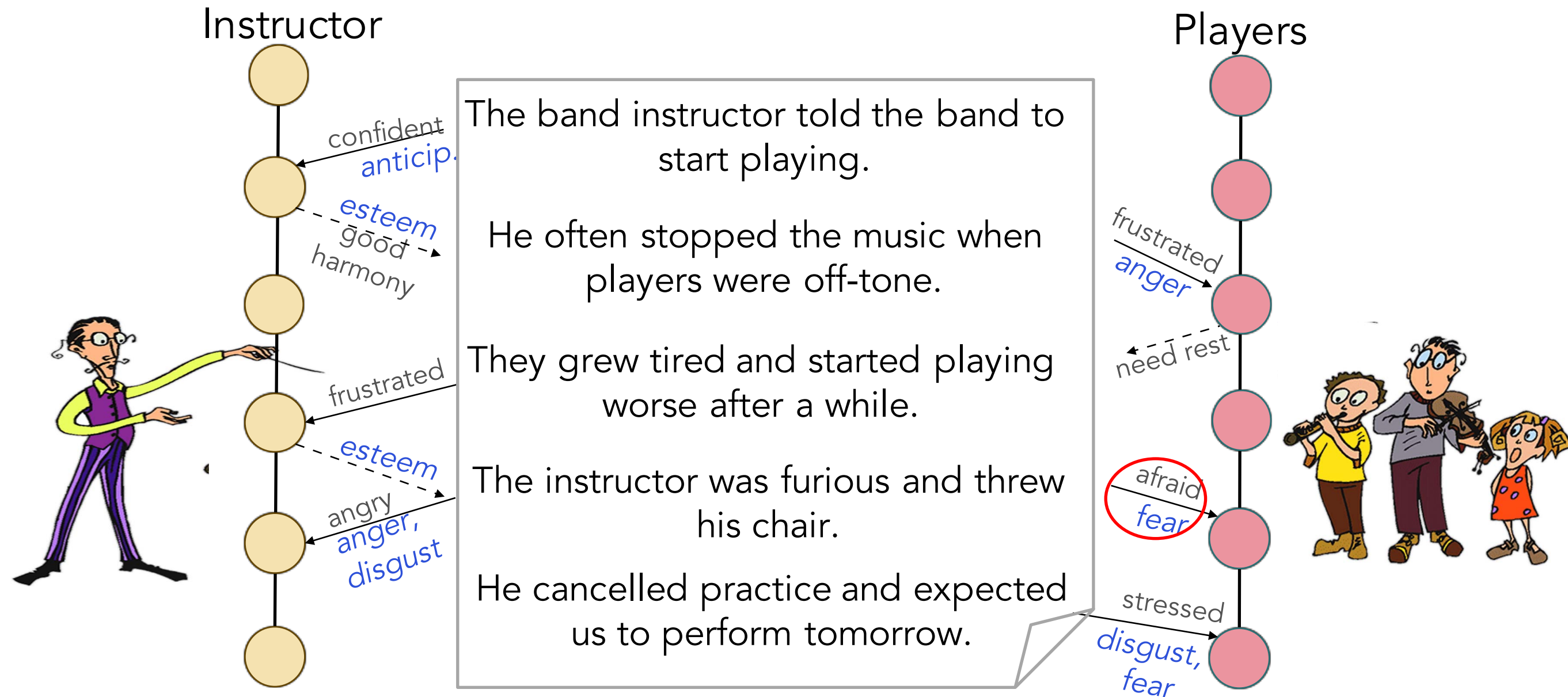


Naïve Psychology of Story Characters (ACL 2018)

Social commonsense

- Script knowledge of events and stories
- Modeling naïve psychology of people
- New challenge datasets
- Unifying representation formalism and models

Reasoning about Naïve Psychology of Story Characters



Commonsense Inference

PersonX cooks thanksgiving dinner

X's intent

to impress their family

X's reaction

tired, feel a sense of belonging

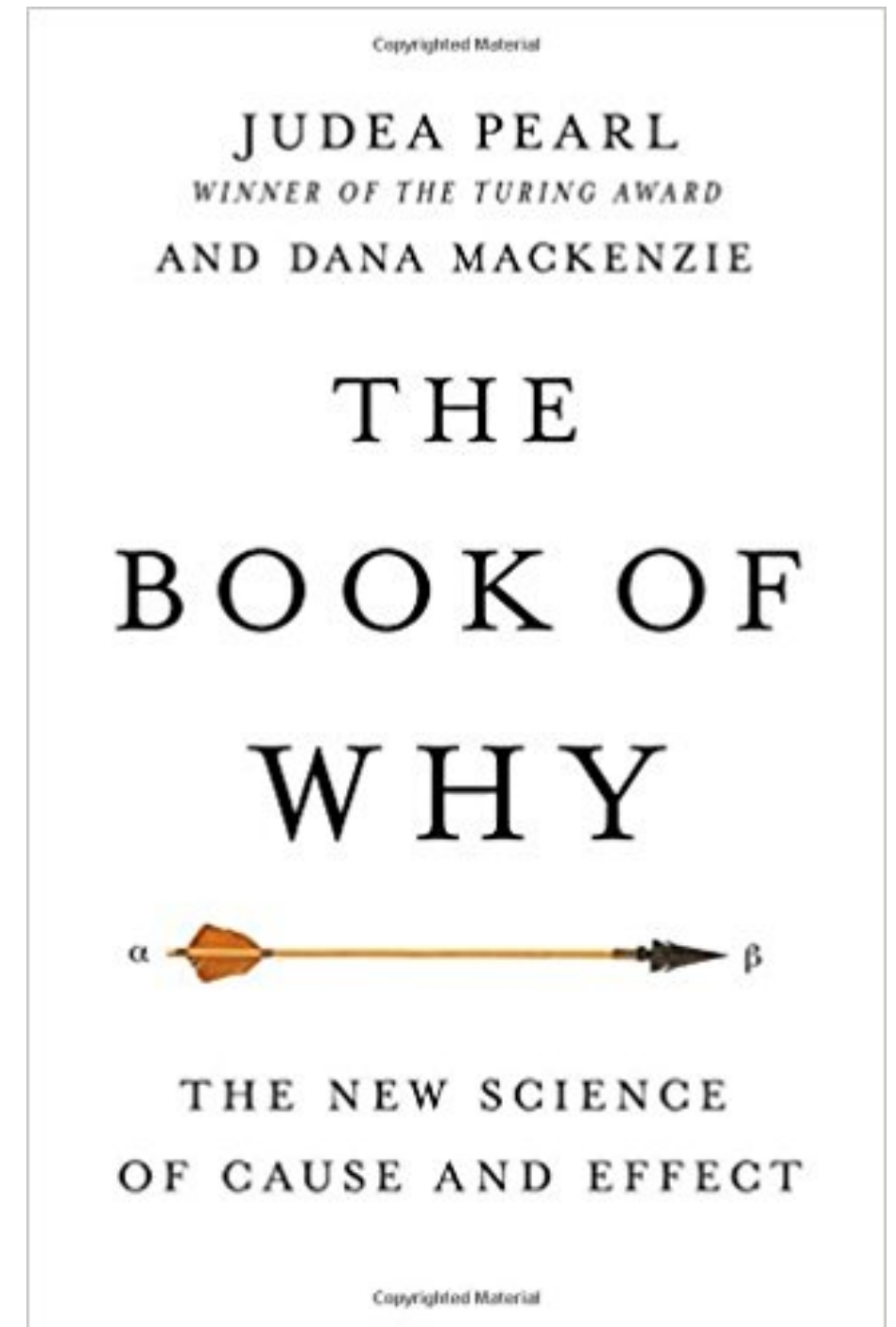
others' reactions

impressed

- Intent
 - mental pre-condition
 - of the agent (X)
- Emotional reactions
 - mental post-condition
 - of the agent (X) and of others (Y) if inferable

"Cause and Effect"

- *"To build truly intelligent machines, teach them cause and effect"* – Pearl, 2018



Simulating Action Dynamics with Neural Process Network

Antoine Bosselut et al. (ICLR 2018)



Need commonsense
to reason about
unseen situations

Coherent Generation Checklist Models

wer"

Ingredients: cauliflower, cooking oil, sauce, salt,

"Are RNNs
a mouth without
a brain?"

Checklist Models
(et al., 2016)

Forgot
to cook
cauliflower!

- not robust
in unfamiliar
situations

Wash and dry the cauliflower.

Heat the oil in a skillet and fry the sauce
until they are golden brown.

Drain on paper towels.

Add the sauce to the sauce and mix well.

Serve hot or cold.

Motivation

- Recurrent Neural Networks (RNNs) are highly effective in learning fluent surface patterns in language
- Without the ability to read between the lines and reason about the unspoken, but obvious facts

"Fry tofu in the pan"

Location of tofu = pan
Temperature of tofu = hot



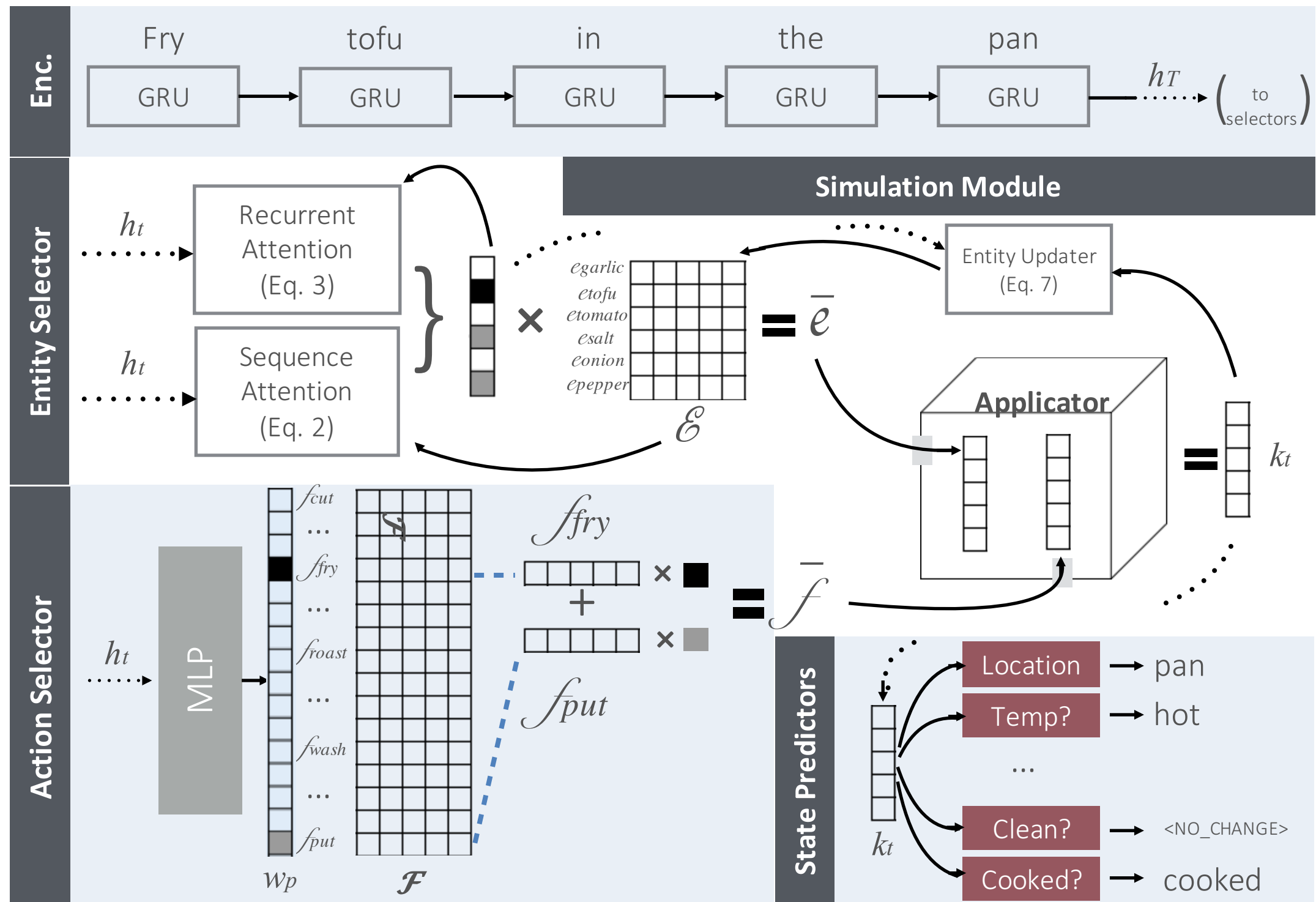
Mental Simulation

- “... the hypothesis that many intuitive physical inferences are based on a *mental physics engine* that is analogous in many ways to the *machine physics engines* used in building interactive video games ...
- “This hypothesis also explains several ‘physics illusions’, and helps to inform the development of artificial intelligence (AI) systems with more *human-like common sense*.”
 - Ullman TD, Spelke E, Battaglia P, Tenenbaum JB (2017)

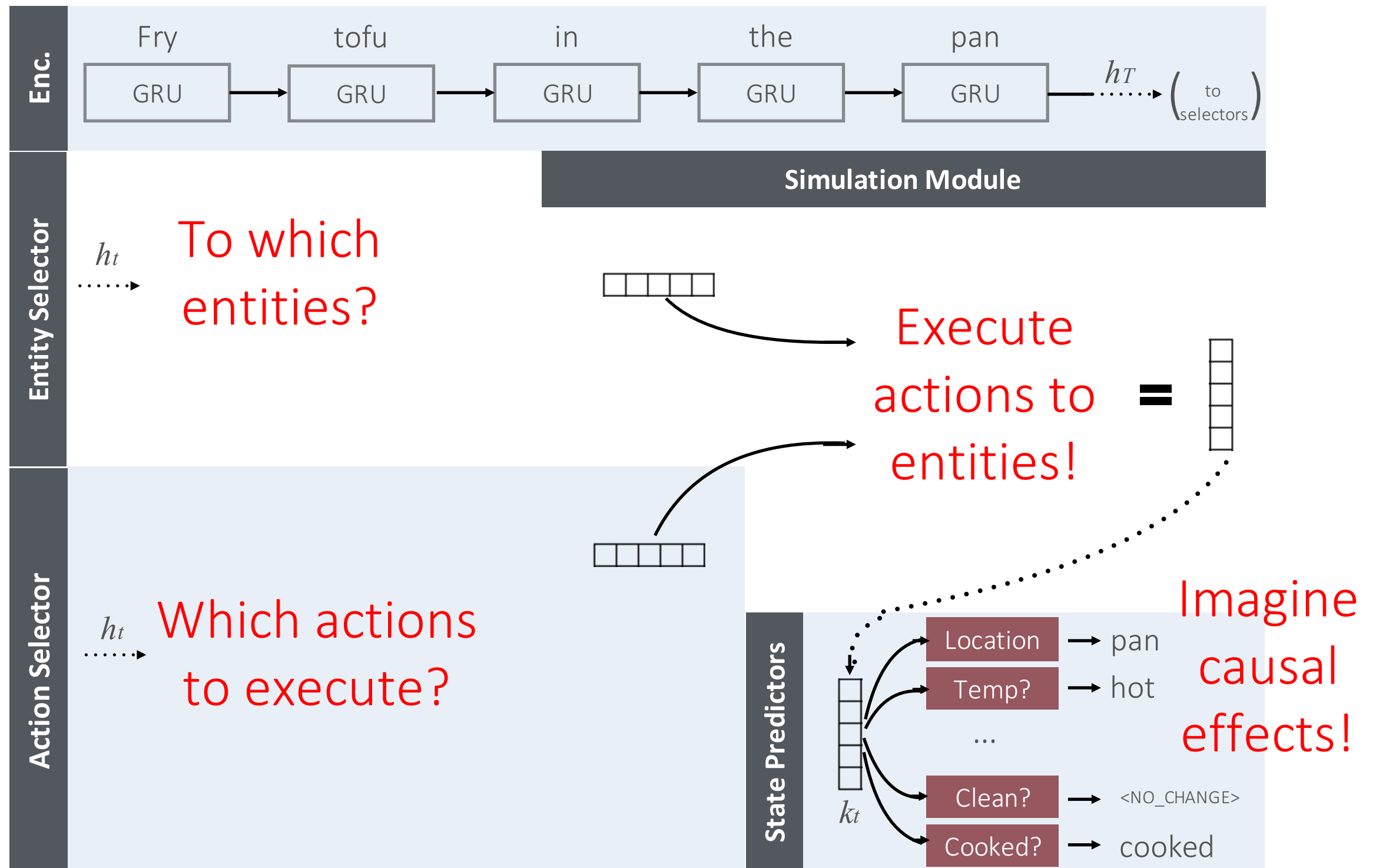
Understanding by Simulation

- Understanding by *Simulation*
 - Simulating the causal effects implied by text
 - Focus on “what is said” + “what is *not* said but implied”
 - Abstracting away from the surface strings
 - (Recurrent Entity Networks (Henaff et al., 2016), Memory Networks (Weston et al., 2015, Sukhbaatar et al., 2016))
- Understanding by *Labeling*
 - Labeling syntactic/semantic categories to surface words
 - Focus on “what is said”
 - Many prior NLU models under this paradigm

Neural Process Networks



Neural Process Networks



Concluding Remarks

- Limitations of NLG point to new challenges of NLU
 - NLU traditionally focuses on understanding only ***natural*** language
 - NLG requires understanding ***machine*** language that is potentially ***unnatural***
- Limitations of LMs
 - While universally useful, LMs are **passive** learners and **surface** learners

Thanks! Questions?

