Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation

Kushal Arora[1], Layla El Asri[2] , Hareesh Bahuleyan[3], Jackie Chi Kit Cheung[1]

1. McGill University/Mila, 2. Borealis AI, 3. Zalando SE

What the heck is exposure bias?

The O.G. definition

“LM training process is very brittle because the model was trained on a different distribution of inputs, namely, words drawn from the data distribution, as opposed to words drawn from the model distribution. As a result the errors made along the way will quickly accumulate. We refer to this discrepancy as exposure bias which occurs when a model is only exposed to the training data distribution, instead of its own predictions.”

- Ranzato et al 2015

Assertion

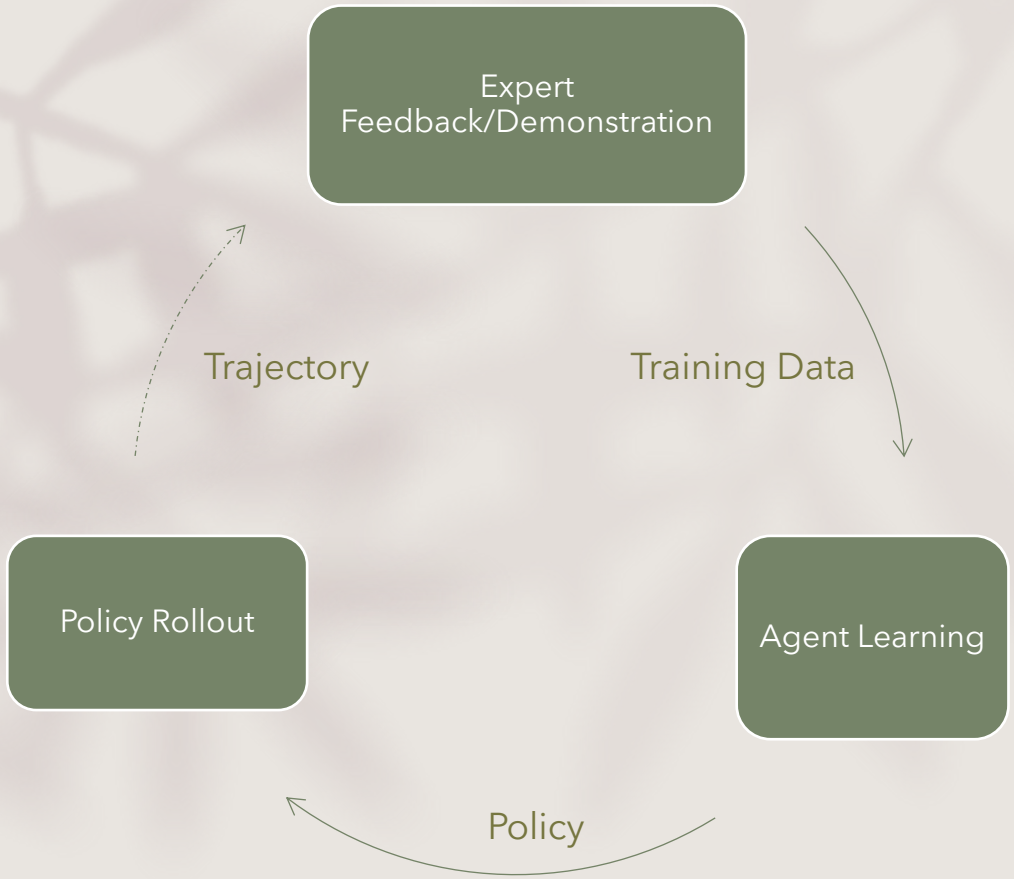
- Higher exposure bias implies worse generation quality.
- Reducing exposure bias might result in more robust language models.

But!...

- Little effort has been paid to formally analyze exposure bias,
- Or directly measure its impact on language generation.

“You can’t improve what you can’t measure.”

- Peter Durker

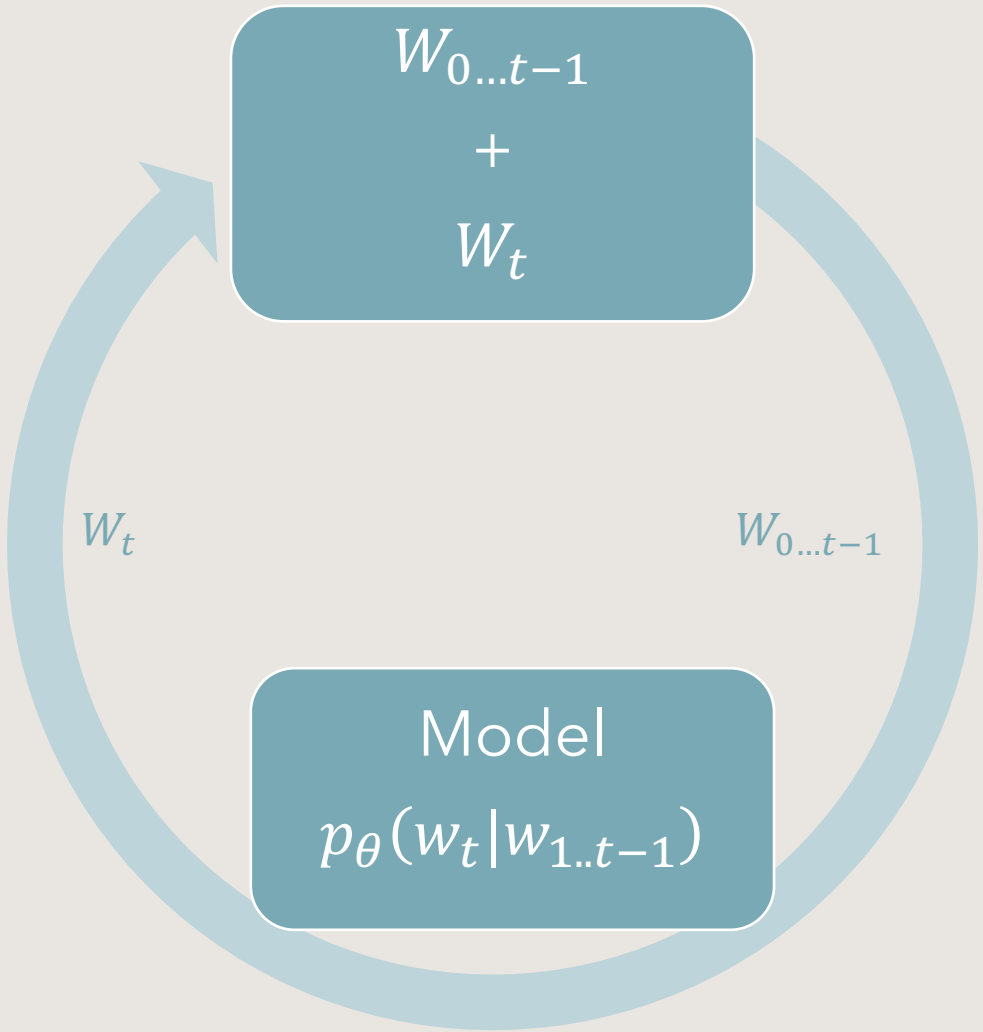


Imitation Learning Paradigm

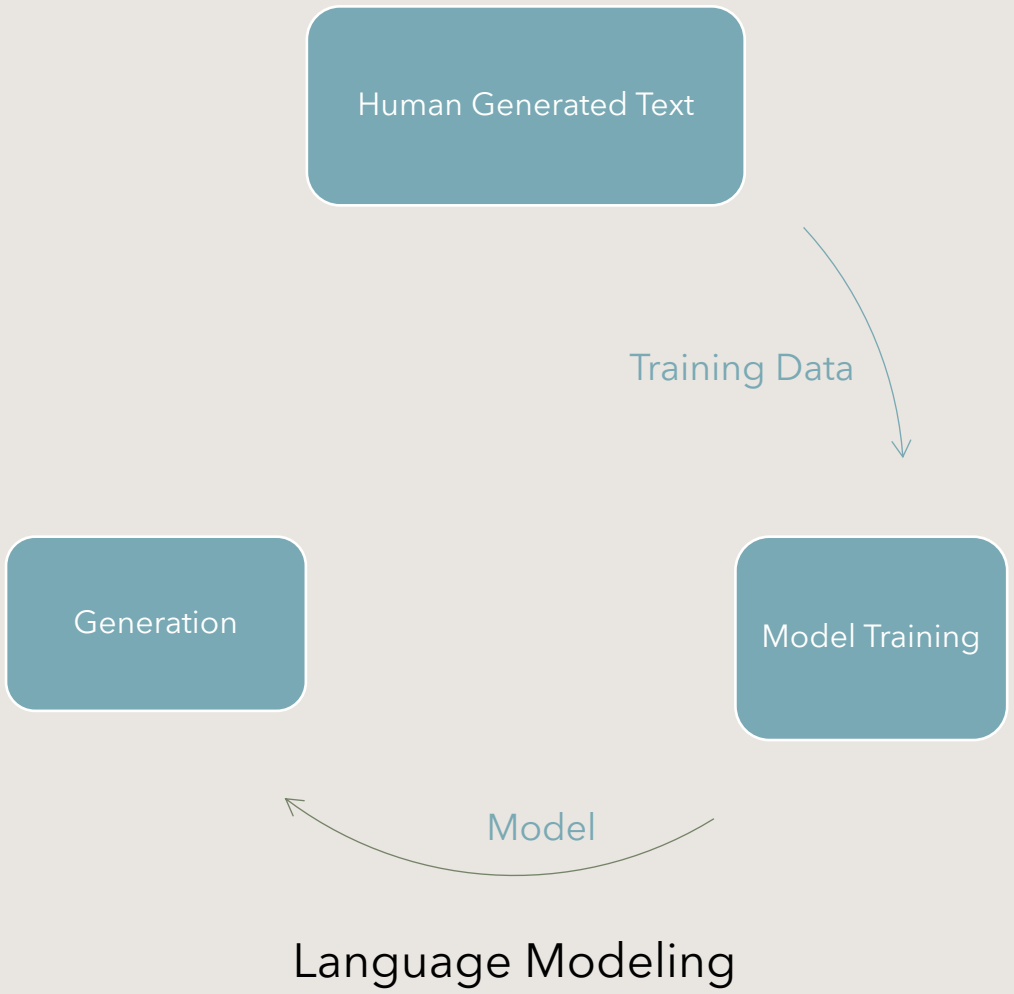
Exposure Bias: An Imitation Learning Perspective

Diagram inspired by: https://smartlabai.medium.com/a-brief-overview-of-imitation-learning-8a8a75c44a9c

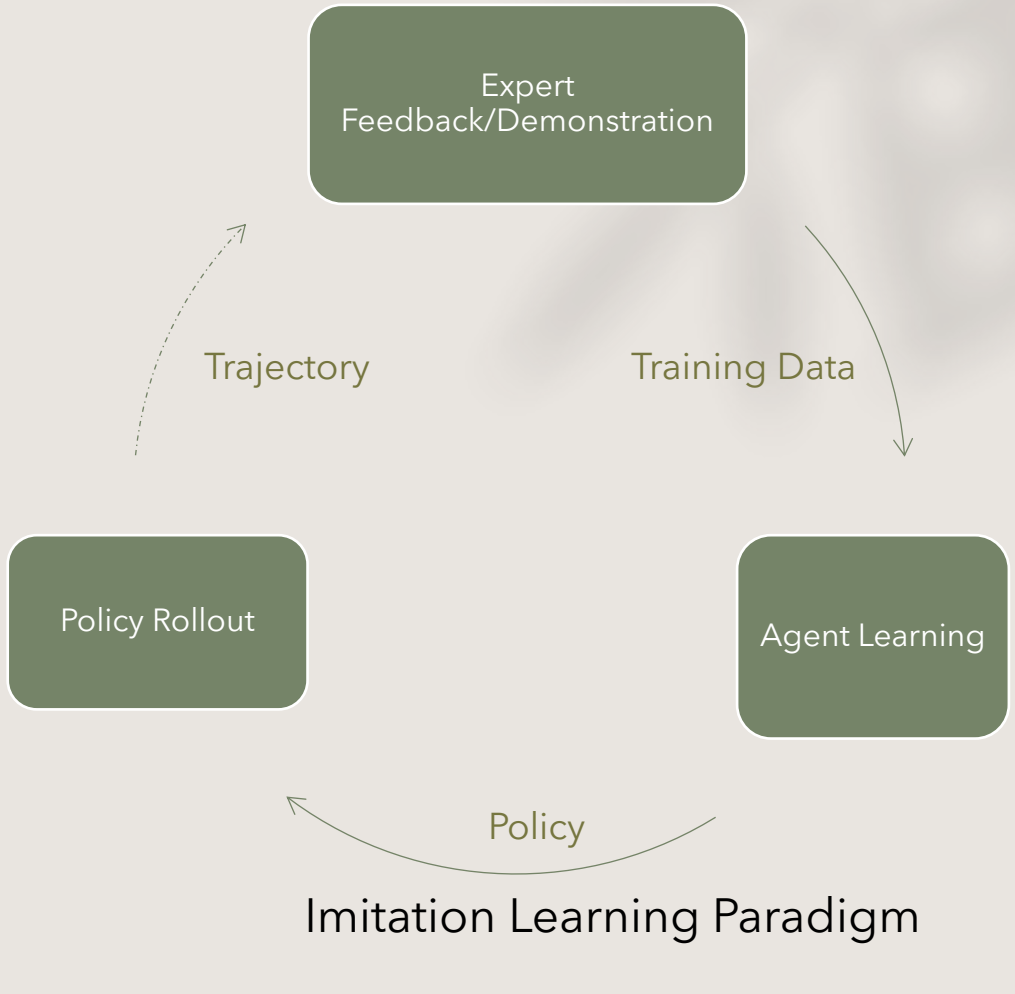
Lang. Generation as a Seq. Decision-Making Process



Language Modeling as Imitation Learning Problem

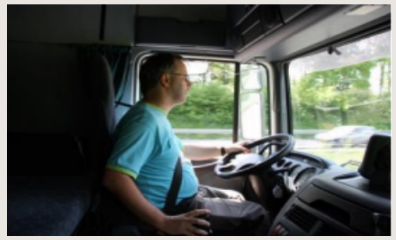


Language Modeling



Imitation Learning Paradigm

MLE Training as Behavior Cloning



Observation Tuple
(s_t, a_t)

Supervised Learning

Policy

Behavior Cloning

$(w_{0\ldots t-1}, w_t)$

MLE Training

Language Model
$p_\theta(w_t|w_{1..t-1})$

Maximum Likelihood Training of Language Models

Behavior Cloning's Error Accumulation Problem

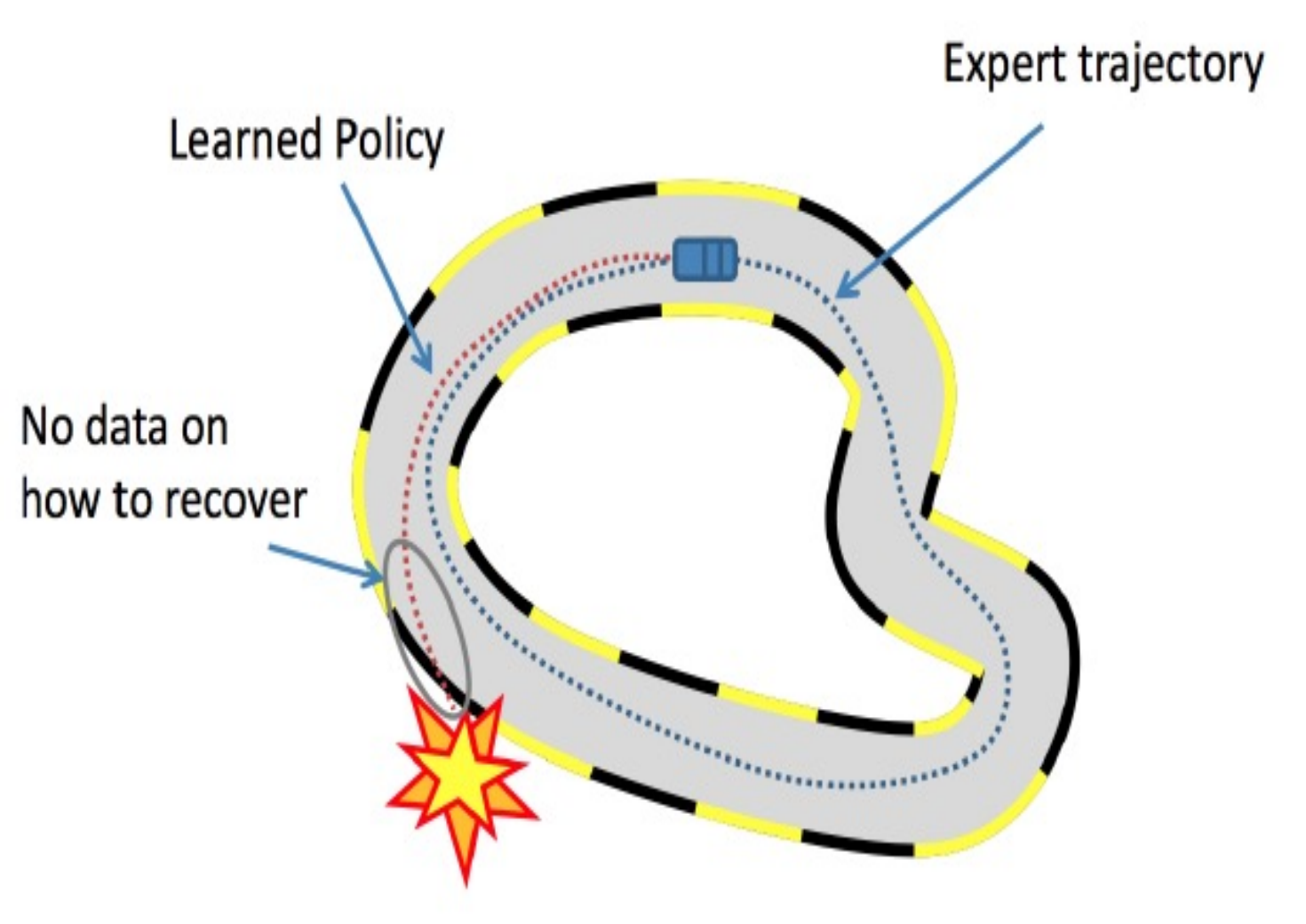


Image credit: https://smartlabai.medium.com/a-brief-overview-of-imitation-learning-8a8a75c44a9c

Behavior Cloning's Error Accumulation Problem

Inference Time Loss of Policy π

Behavior Cloning Policy

Oracle Policy

$$T\epsilon \leq L^I(\pi_{BC}) - L^I(\pi_o) \leq T^2\epsilon$$ [1]

$$R(\pi_{BC})$$

Max. Time Step

Per Step Loss Modeling Error

Inference-time regret of policy π_{BC}

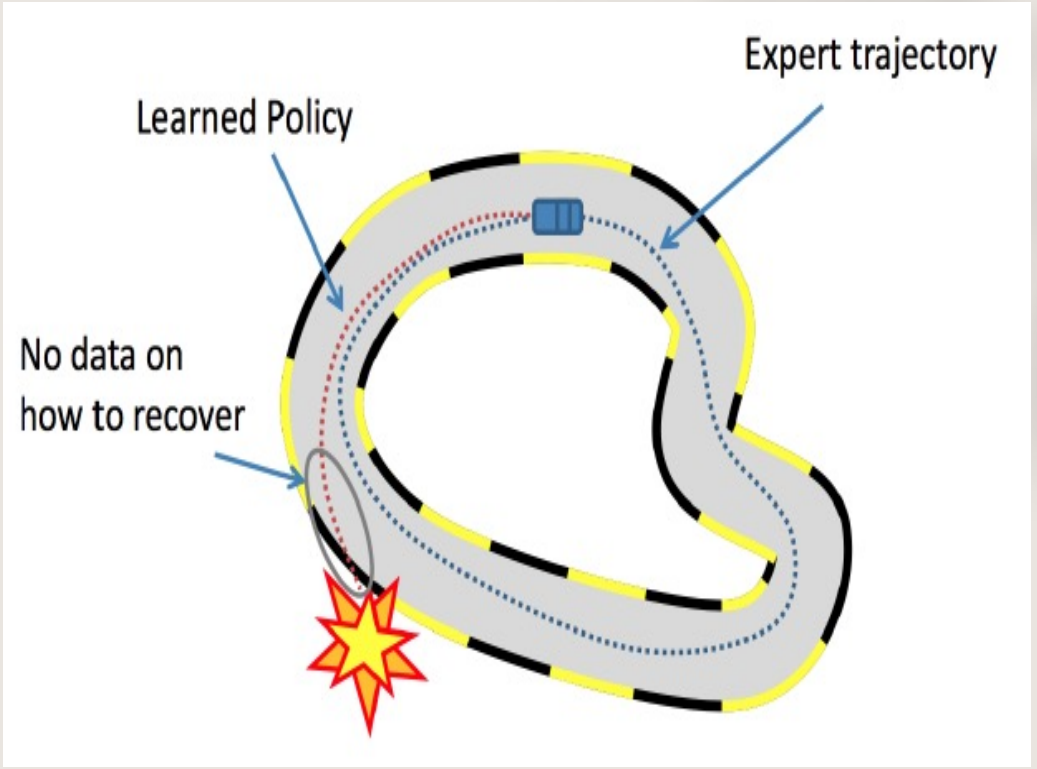


Image credit: https://smartlabai.medium.com/a-brief-overview-of-imitation-learning-8a8a75c44a9c

[1] Ross, Stephane, and Drew Bagnell. "Efficient Reductions for Imitation Learning."

Exposure Bias and Error Accumulation

Till Now,

- We formulated language generation as sequential decision-making problem
- Showed MLE-based training is equivalent to behavior cloning*.
- Discussed the error accumulation issue in behavior cloning, and
- Presented an analytical framework from IL literature to analyze this error accumulation.

This helps us analyze exposure bias as error accumulation due to reduction of a seq. decision-making problem to a supervised learning problem.

* under a particular choice of loss function. See papers for more details.

Quantifying Error Accum. Due to Exposure Bias

- Exploiting the behavior cloning and MLE training equivalence,
- We borrow the regret-based error analysis for IL literature.
- We define accumulation of error $AccError_{\leq}(l)$ as

$$AccError_{\leq}(l) = R_{\leq l}(p_\theta, F) / \epsilon_{\leq l}$$

Inference-time regret of model p_θ and decoding algorithm F

Per-step error/model error of language model p_θ

Does Exposure Bias lead to Accumulation of Errors?

From behavior cloning's error accumulation analysis, we can bound $AccError_{\leq}(l)$ as:

$$AccError_{\leq}(l) = R_{\leq l}(p_\theta, F)/\epsilon_{\leq l}$$

Inference-time regret of model p_θ and decoding algorithm F

Per-step error/model error of language model p_θ

$$l \leq AccError_{\leq}(l) \leq l^2$$

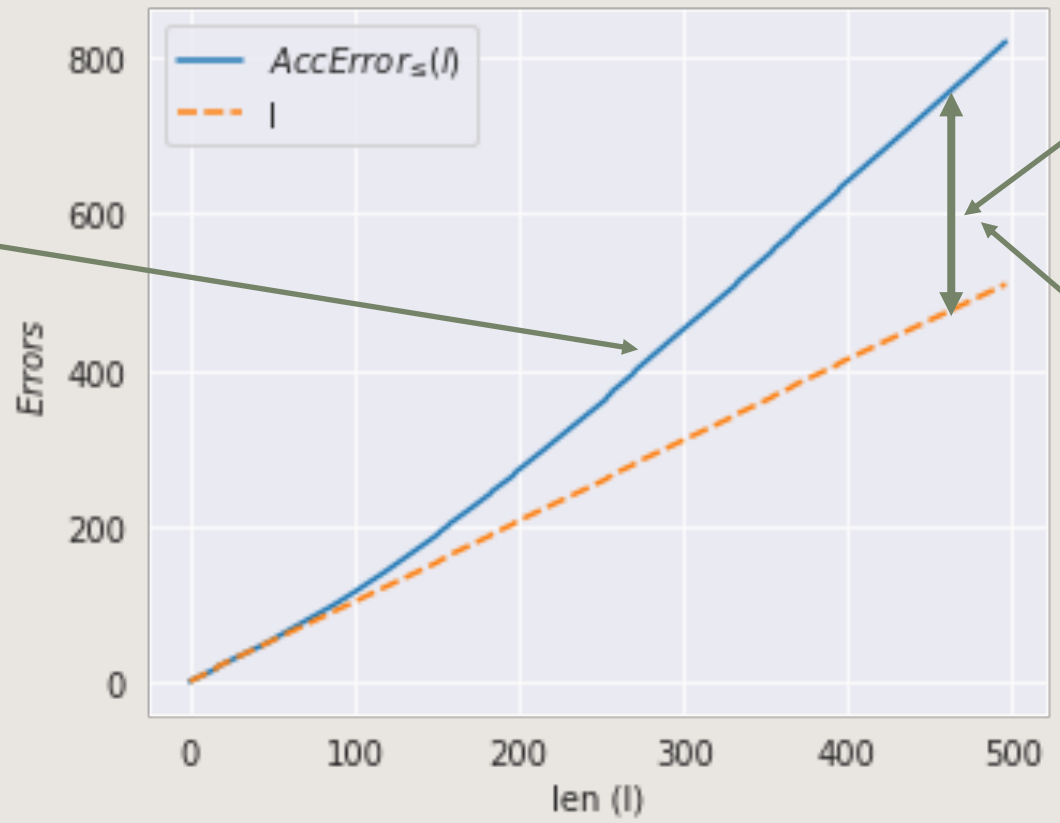
Min. #errors if there is no error accumulation due to exposure bias.

Super-linear growth indicates error accumulation due to exposure bias.

Max. #errors if there is no error accumulation due to exposure bias.

Does Exposure Bias lead to Accumulation of Errors?

Accumulated Errors grow super-linearly with the sequence length.



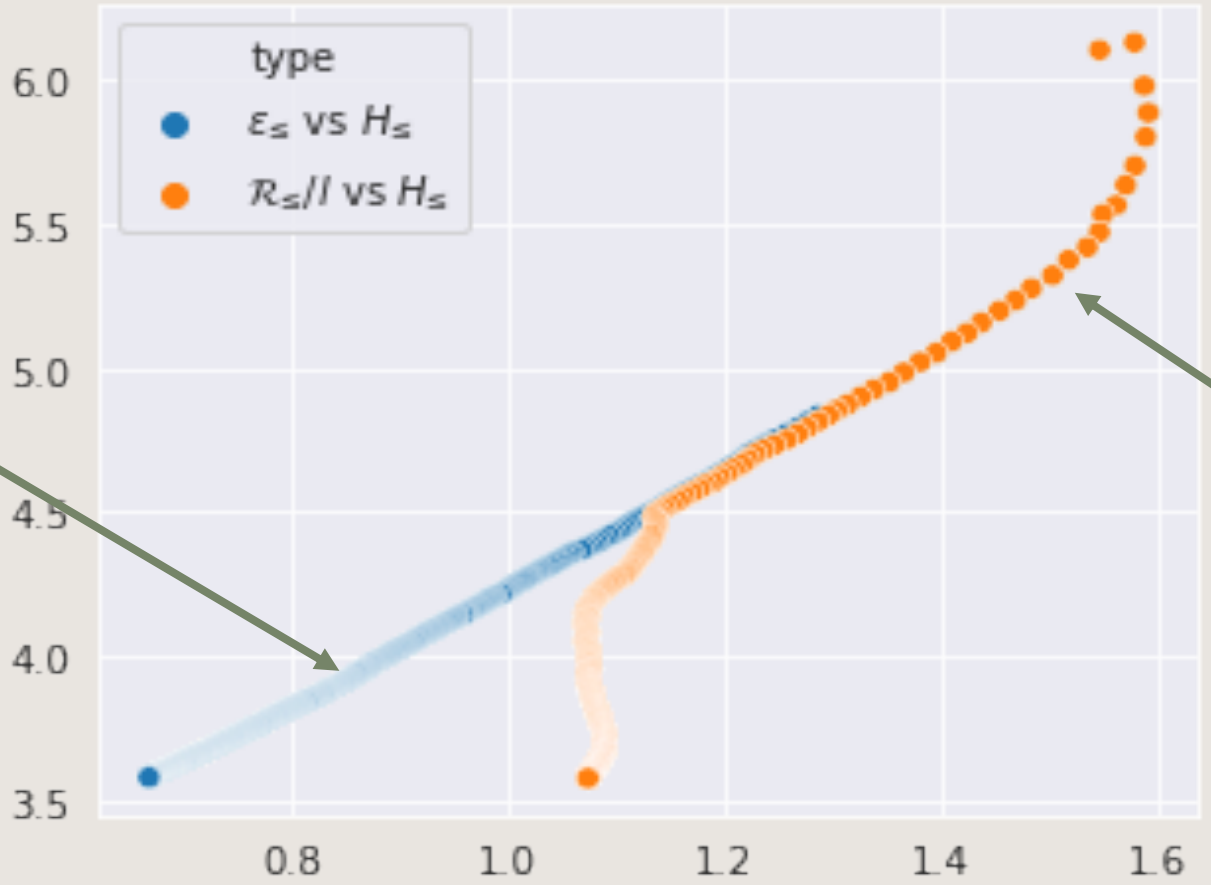
Excess Error due to Exposure Bias

Excess Error increases monotonically with the sequence length.

Exposure bias does lead to error accumulation!

Why Perplexity is Not Enough

Near Perfect Correlation (ρ=0.9997) between per-step error (ε) and cross entropy (H)



Weak Correlation (ρ=0.4003) between accumulated error ($AccError$) and cross entropy (H)

Perplexity fails to capture the accumulation due to exposure bias!

Error Accumulation Impacts Generation Quality

Search	%ExErrAcc (↓)	Generation Quality			
		seq-rep-4 (↓)	rep (↓)	wrep (↓)	uniq (↑)
Greedy	60.96%	0.8990	0.4423	0.4136	7833
Beam (k=5)	69.72%	0.8094	0.4064	0.3787	10966
Sampling					
w/ Temp (temp=1)	39.37%	0.1883	0.2547	0.2301	23729
w/ Temp (temp=1.2)	**24.75%**	**0.1556**	**0.2271**	**0.2033**	25225
w/ top-k (k=100)	35.37%	0.1690	0.2409	0.2166	**26251**
w/ top-p (p=0.94)	48.71%	0.2218	0.2743	0.2490	22582
Human	-	0.0274	0.4338	-	28739

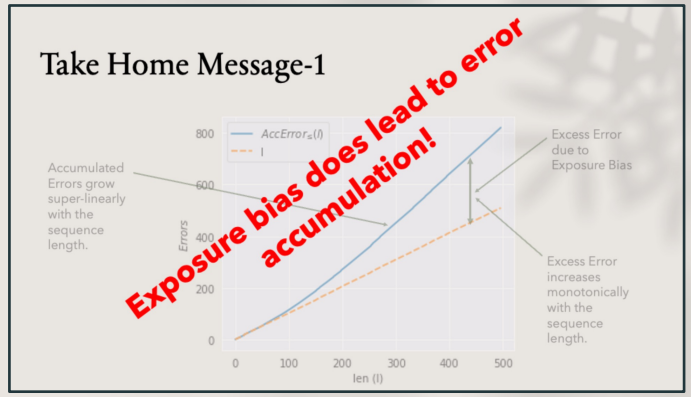
Exposure Bias ↓

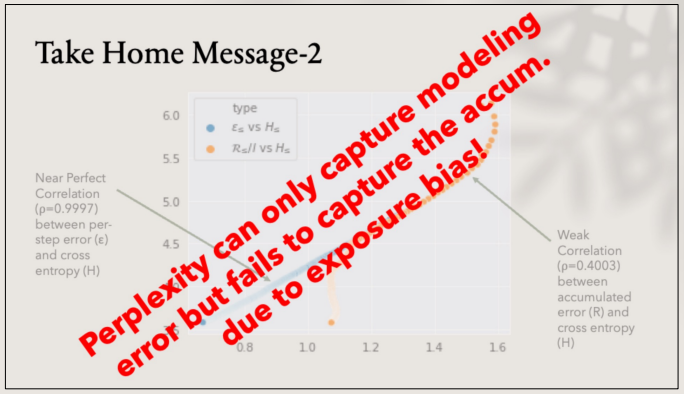
Generation Quality ↑

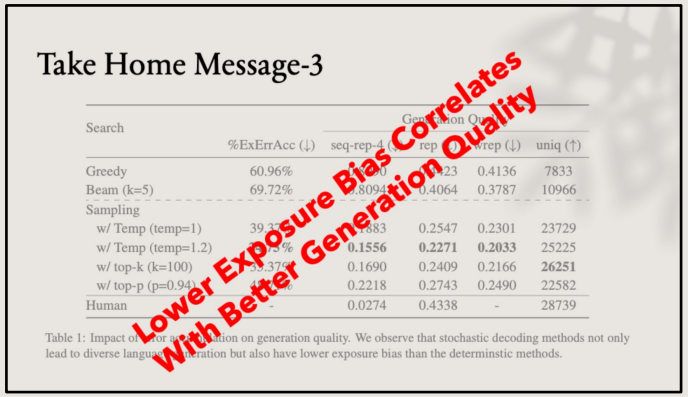
Lower exposure bias correlates with better generation quality

Take-Home Messages

- Exposure bias does lead to error accumulation!
- Perplexity fails to capture the accumulation due to exposure bias!
- Lower exposure bias correlates with better generation quality!







Thanks

- For details, please see our paper: https://arxiv.org/abs/2204.01171
- Code is available at: https://github.com/kushalarora/quantifying_exposure_bias
- Please visit us at our poster at ACL 2022.