Text Summarization beyond Seq2Seq Models for Salience, Faithfulness, and Factuality

Yue Dong

PhD thesis defense

Reasoning & Learning Lab, School of Computer Science, McGill University Montreal Institute for Learning Algorithms (MILA)

Monday Nov. 28th, 2022

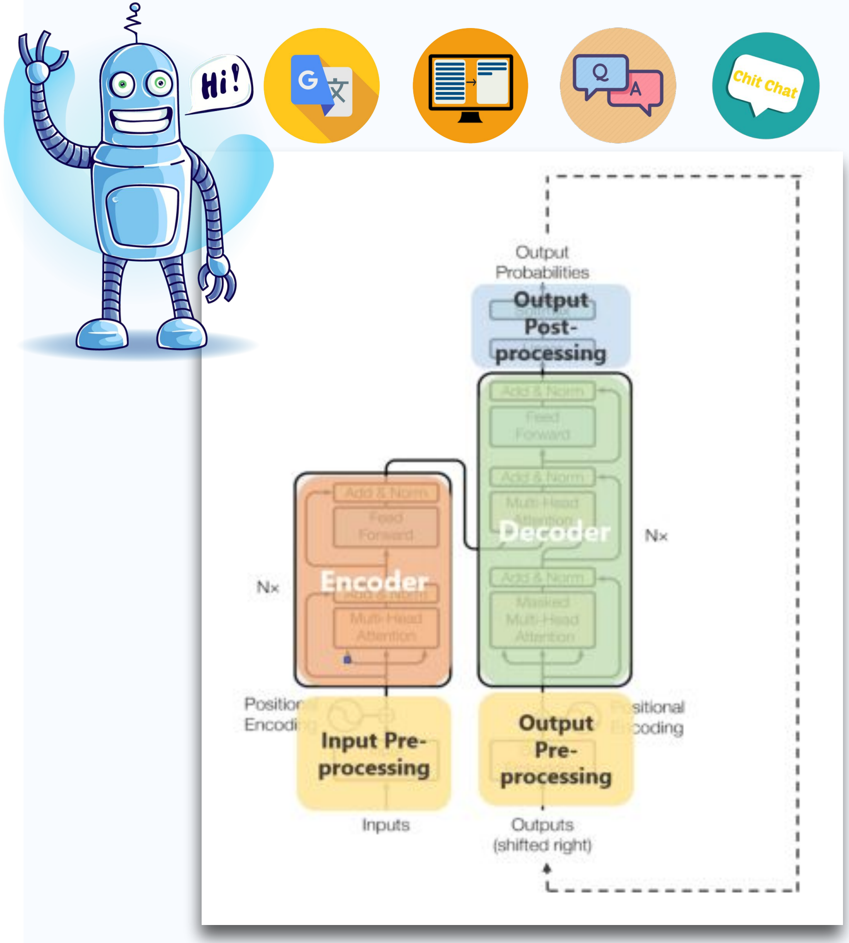
Natural Language Generation (NLG)

Goal:
generate human-like language with context/condition

Tasks:
machine translation, Q&A, summarization, dialogue etc

Dominant models:
sequence-to-sequence models

- text-based
- encoder-decoder architecture
- autoregressive

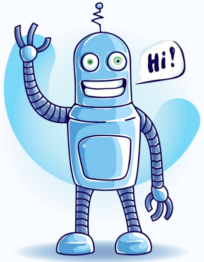


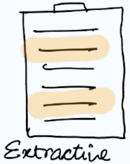
Thesis Scope:
Text Summarization

Shortening text while preserving **main ideas**

Source

A fire crew remains at Plasgran, Wimblington. The incident began more than 16 hours ago. Road closures are expected …





A fire crew remains at Plasgran, Wimblington.

A large **fire** has broken out **at Plasgran in** Cambridgeshire.

Summarization Requirements

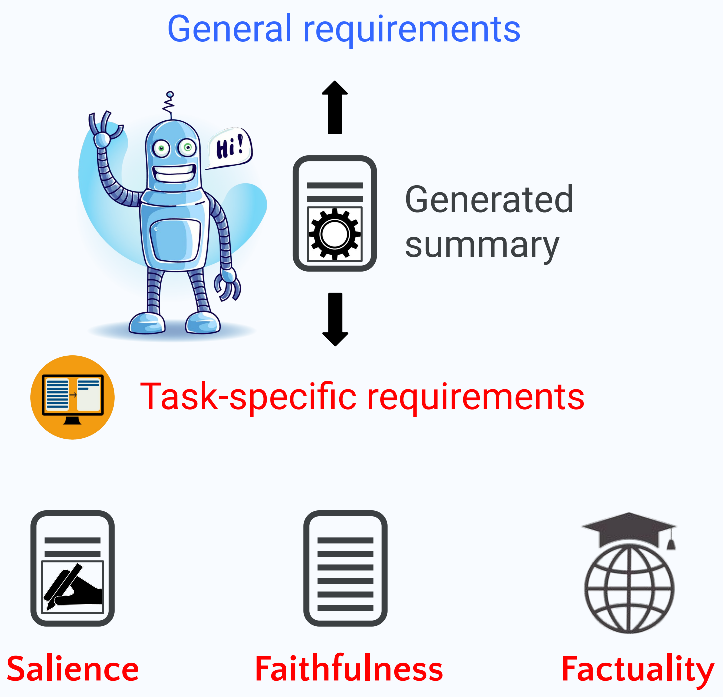
A good system summary should be:

a. Fluent
b. Natural (human-like)

a. Salient
   - contain **important key points**

b. Faithful
   - consistent with **the source**

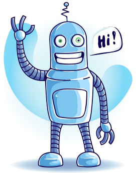
c. Factual
   - consistent with the **world knowledge**



Kedzie, Christopher. *Salience Estimation and Faithful Generation: Modeling Methods for Text Summarization and Generation*. Columbia University, 2021.

Trends in NLG: Go Generic and Go Big

Impressive human-like (natural and fluent) generations [1]



Learn task-specific requirements **implicitly**

- Data-intensive
- Hard to control
- **Reliability?**

[1] Kaplan et al., *Scaling laws for Neural Language Models*, 2021

Seq2seq	Seq2seq wo. attention	Seq2seq w. Attention	Transformer
Encoder	RNN/CNN	RNN/CNN	attention
Decoder	RNN/CNN	RNN/CNN	attention
Decoder-encoder interaction	static fixed-sized vector	attention	attention

Less inductive bias →

Less task-specific focus

Paradigm	Supervised learning	Transfer learning	Prompt-based learning
Generic pre-training	×	✓	✓
Task adaptation	training from scratch [2]	task specific fine-tuning [3]	multitask instruction-tuning [4]

[2] Sutskever et al., *Sequence to sequence learning with neural networks.* NeurIPS 2014
[3] Raffel et al. *Exploring the limits of transfer learning with a unified text-to-text transformer.* JML 2020.
[4] Sanh et al. *Multitask Prompted Training Enables Zero-Shot Task Generalization.* ICLR 2022

Thesis Statement

Designing models with **appropriate inductive bias** beyond the standard seq2seq setu is effective to meet requirements **specific** to text summarization

Inductive bias in modeling employs prior knowledge to determine a learner’s hypothesis space

Salience

Seq2Set - control bias exploitation

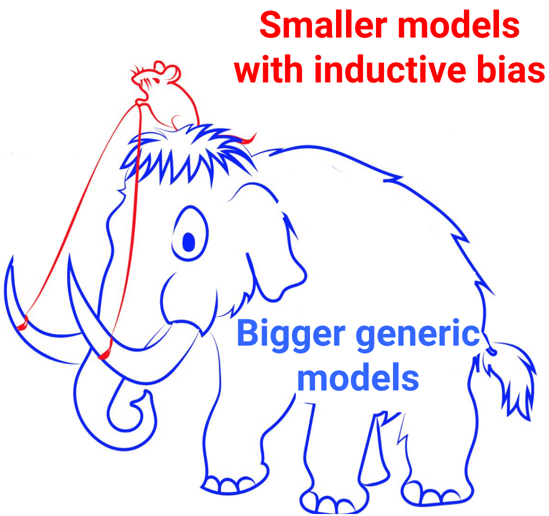
Faithfulness

Seq2Edit - control hallucination

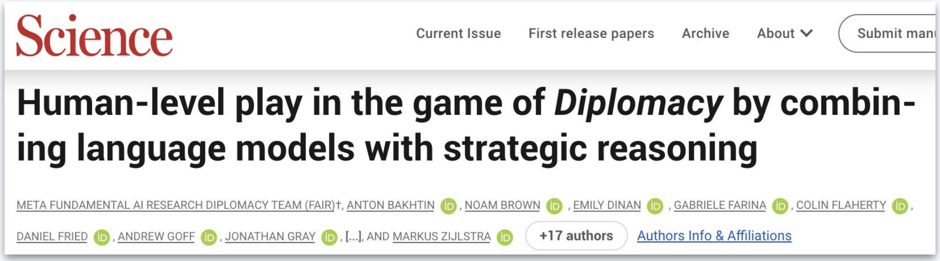
Factuality

Seq + KB - control with facts

Cooperation: Go big and Go Under Control



Adapted from PPLM, Dathathri and Madotto et al., ICLR 2020 (GitHub)



Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning

META FUNDAMENTAL AI RESEARCH DIPLOMACY TEAM (FAIR)†, ANTON BAKHTIN, NOAM BROWN, EMILY DINAN, GABRIELE FARINA, COLIN FLAHERTY, DANIEL FRIED, ANDREW GOFF, JONATHAN GRAY, [...], AND MARKUS ZIJLSTRA +17 authors Authors Info & Affiliations

Cicero

ranked in the top 10% of human participants

- **Dialogue model base:**
  - 2.7B BART
- **Many inductive biases**:
  - Controlling natural language generation via planning, RL, neuro-symbolic KB, filter, and ranker, etc.

https://ai.facebook.com/blog/cicero-ai-negotiates-persuades-and-cooperates-with-people/, Nov. 22, 2022

BanditSum

Extractive Summarization as a Contextual Bandit

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, Jackie Chi Kit Cheung

EMNLP 2018
Oral

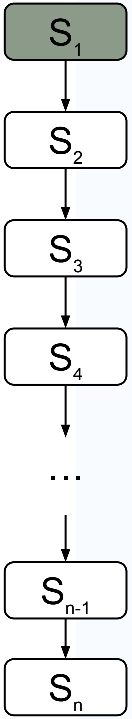
Control bias exploitation with non-autoregressive models

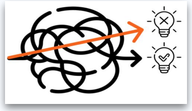
Salience in Extractive Summarization

Goal: pick a set of salient sentences

Adaptation from seq2seq setting: **sequential** binary labeling

- Exposure bias
- Approximated binary labels
- Prone to exploit **lead bias**



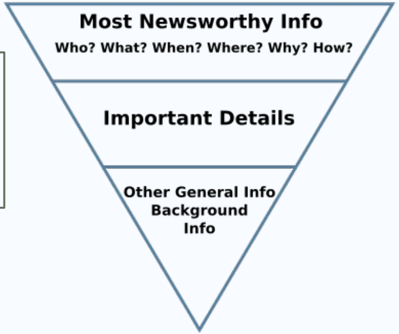


Artifacts & Biases
Always picks the 1st sentence

⬆

Chooses the 1st sentence

⬇



Contents
1st sentence is important in this example

Contextual Multi-armed Bandit

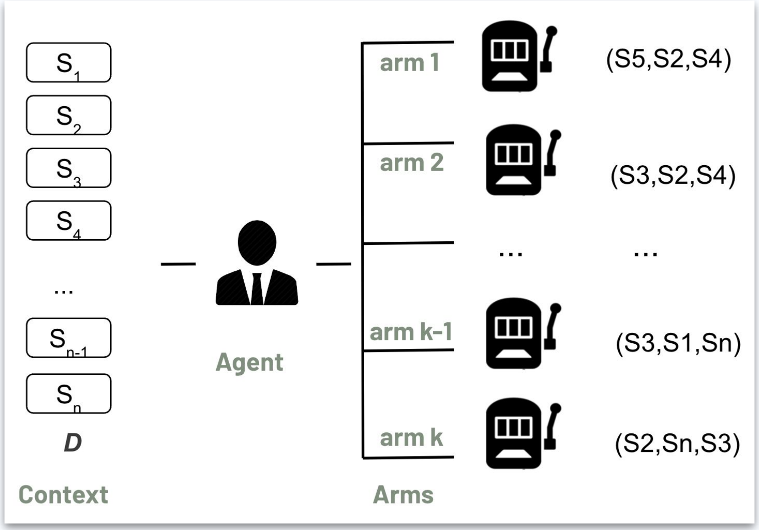
Control bias exploitation with **non-autoregressive models**

- Directly optimize **content importance**
- Trained by REINFORCE
- Selection **regardless of position** in the document

Context = the document

Arm = a set of *M* sentences

Reward = f (arm, context)



BanditSum: RL in a Nutshell

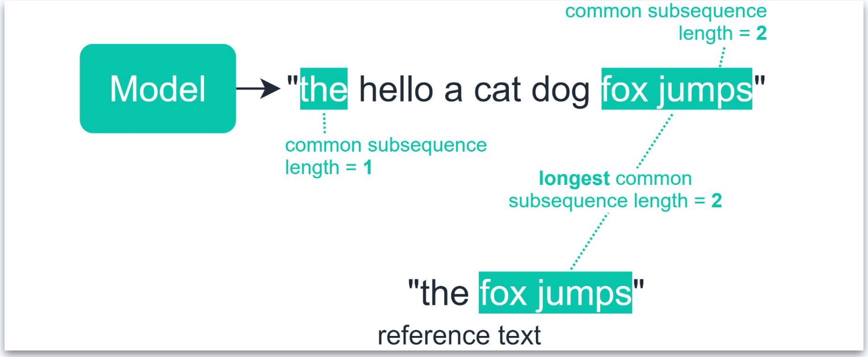
Goal: generate **a summary i that maximize reward R**, based on the **reference summary a**

$$J(\theta) = E\left[R(i,a)\right] \qquad (1)$$

Policy gradient reinforcement learning likelihood ratio gradient estimator (Williams, 1992)

$$\nabla_\theta J(\theta) = E\left[\nabla_\theta \log p_\theta(i|d) R(i,a)\right] \qquad (2)$$

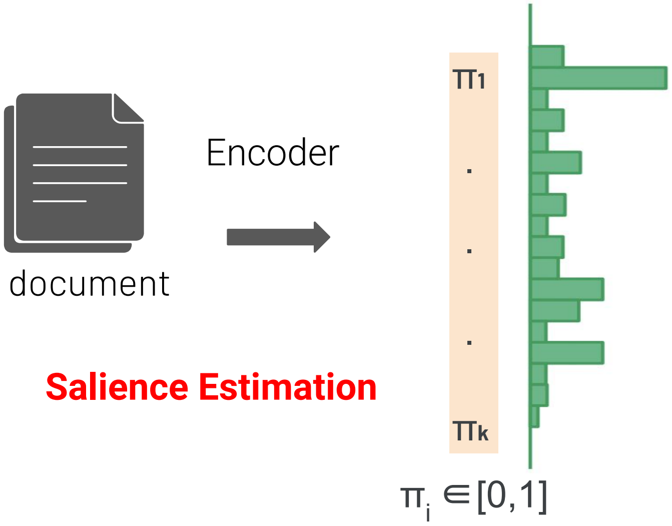
ROUGE: similarity between generated summary and gold-reference summary


common subsequence length = 2
Model
"the hello a cat dog fox jumps"
common subsequence length = 1
longest common subsequence length = 2
"the fox jumps"
reference text

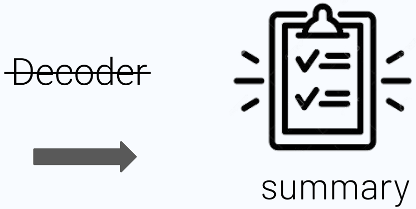
$$R(i,a) = \frac{1}{3} \sum_{k=1,2,L} \text{ROUGE-}k_f(i,a)$$

Structure of Policy $p_\theta(\cdot|d) = \mu(\cdot|\pi_\theta(d))$

Deterministic $\pi_\theta(d)$



Stochastic $p_\theta(i|d) = \mu(i|\pi_\theta(d))$



Sampling wo. replacement

$$\prod_{j=1}^{M}\left(\frac{\epsilon}{N_d - j + 1} + \frac{(1-\epsilon)\pi(d)_{i_j}}{z(d) - \sum_{k=1}^{j-1}\pi(d)_{i_k}}\right)$$

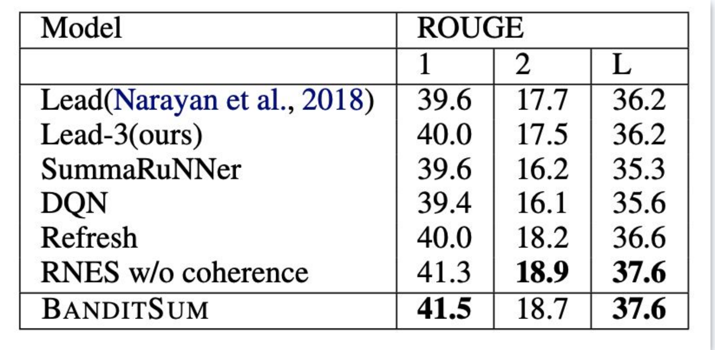
Explore Exploit

Results: Overall

Dataset: CNN/DailyMail

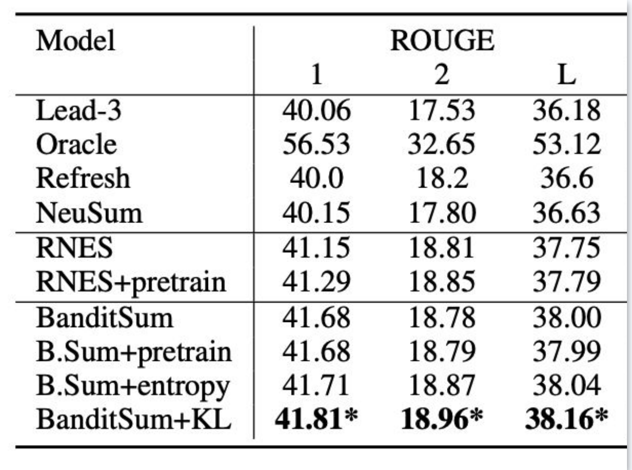
- Outperform seq2seq [1]:
    - ROUGE - 1,2,L + 1.9, 2.5, 2.3
    - Prefered by human judges

- Comparable to seq2seq + RL [2]



Model	ROUGE		
	1	2	L
Lead(Narayan et al., 2018)	39.6	17.7	36.2
Lead-3(ours)	40.0	17.5	36.2
SummaRuNNer	39.6	16.2	35.3
DQN	39.4	16.1	35.6
Refresh	40.0	18.2	36.6
RNES w/o coherence	41.3	**18.9**	**37.6**
BanditSum	**41.5**	18.7	**37.6**

Test results after 2 epochs



Model	ROUGE		
	1	2	L
Lead-3	40.06	17.53	36.18
Oracle	56.53	32.65	53.12
Refresh	40.0	18.2	36.6
NeuSum	40.15	17.80	36.63
RNES	41.15	18.81	37.75
RNES+pretrain	41.29	18.85	37.79
BanditSum	41.68	18.78	38.00
B.Sum+pretrain	41.68	18.79	37.99
B.Sum+entropy	41.71	18.87	38.04
BanditSum+KL	**41.81***	**18.96***	**38.16***

Test results after 4 epochs [3]

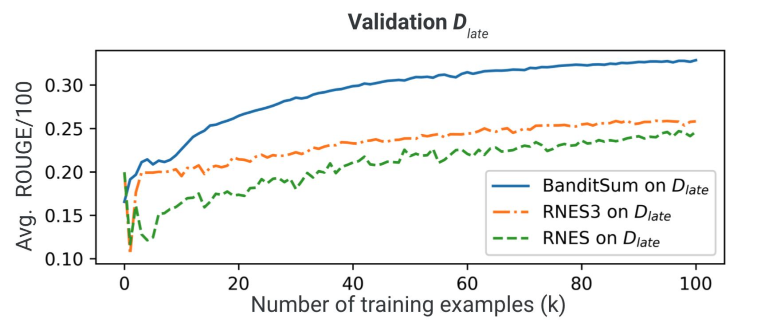
[1] Nallapati et al., *Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.* AAAI 2017.
[2] Wu and Hu. *Learning to extract coherent summary via deep reinforcement learning.* AAAI 2018.
[3] Grenander et al., *Countering the Effects of Lead Bias in News Summarization via Multi-Stage Training and Auxiliary Losses*. EMNLP 2019.

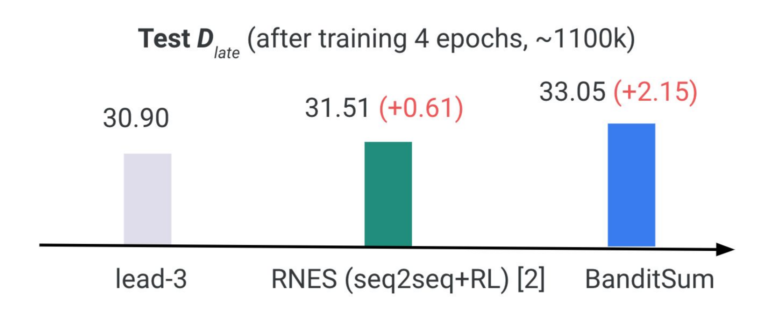
Results: Exploit Less Lead Bias

D_{late} : documents w. salient sentences appear late

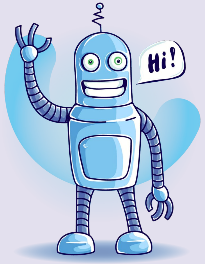
Robust in domain shift compared to seq2seq + RL [2]:

- Sample efficient
- Converge faster





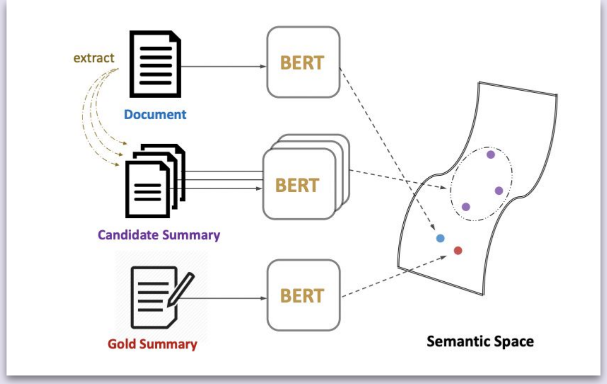
Key Takeaways



- Inductive bias in modeling (e.g., extractive seq2seq) that **coincide with artifacts** (e.g., lead bias) may be the bottleneck to robust generalization

- For extractive summarization, **inductive biases that select sentences regardless of position** for global salience estimation may be promising

Impact: the SOTA model MatchSUM (Zhong et al., 2020) learn to rank combinatorial set of sentences



EditNTS

An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, Jackie Chi Kit Cheung

ACL 2019
Oral

Control hallucination via edits

Hallucination

Hallucination: generate[d] text that is nonsensical, or inconsistent **with the provided input**

Causes [1]:

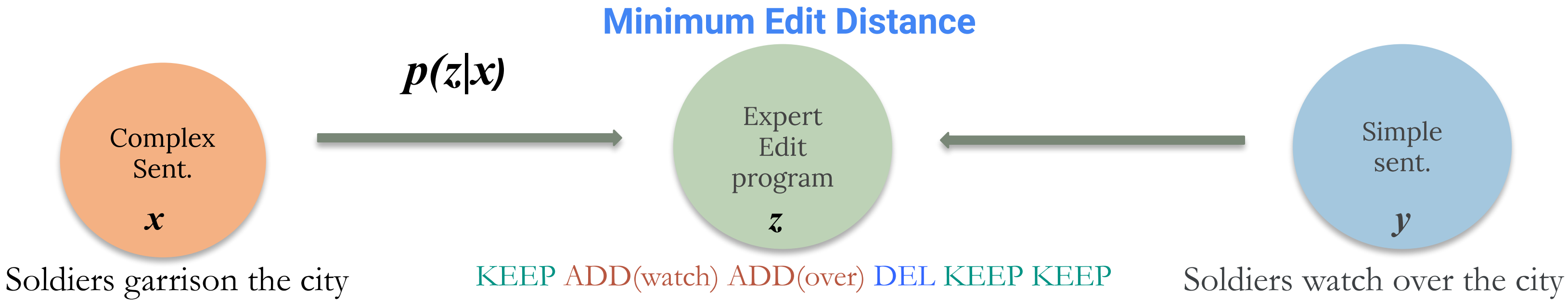
1. **Divergence of source texts and references** in training data
2. **Memorized (factual) knowledge** in models with a really high parameter count (e.g., T5-11B)
3. In general, **model quality** issues

[1] Ji, Ziwei, et al. *Survey of hallucination in natural language generation.* ACM Computing Surveys 2022.

Control Hallucination by Editing Inputs

Our proposal (Seq2Edit):

- Bounds the generation freedom by learning edits
- **Generates** natural language by applying **edit operations** to the **input text**



Minimum Edit Distance

$p(z|x)$

Complex Sent. x

Expert Edit program z

Simple sent. y

Soldiers garrison the city

KEEP ADD(watch) ADD(over) DEL KEEP KEEP

Soldiers watch over the city

EditNTS: Edit-based Learning

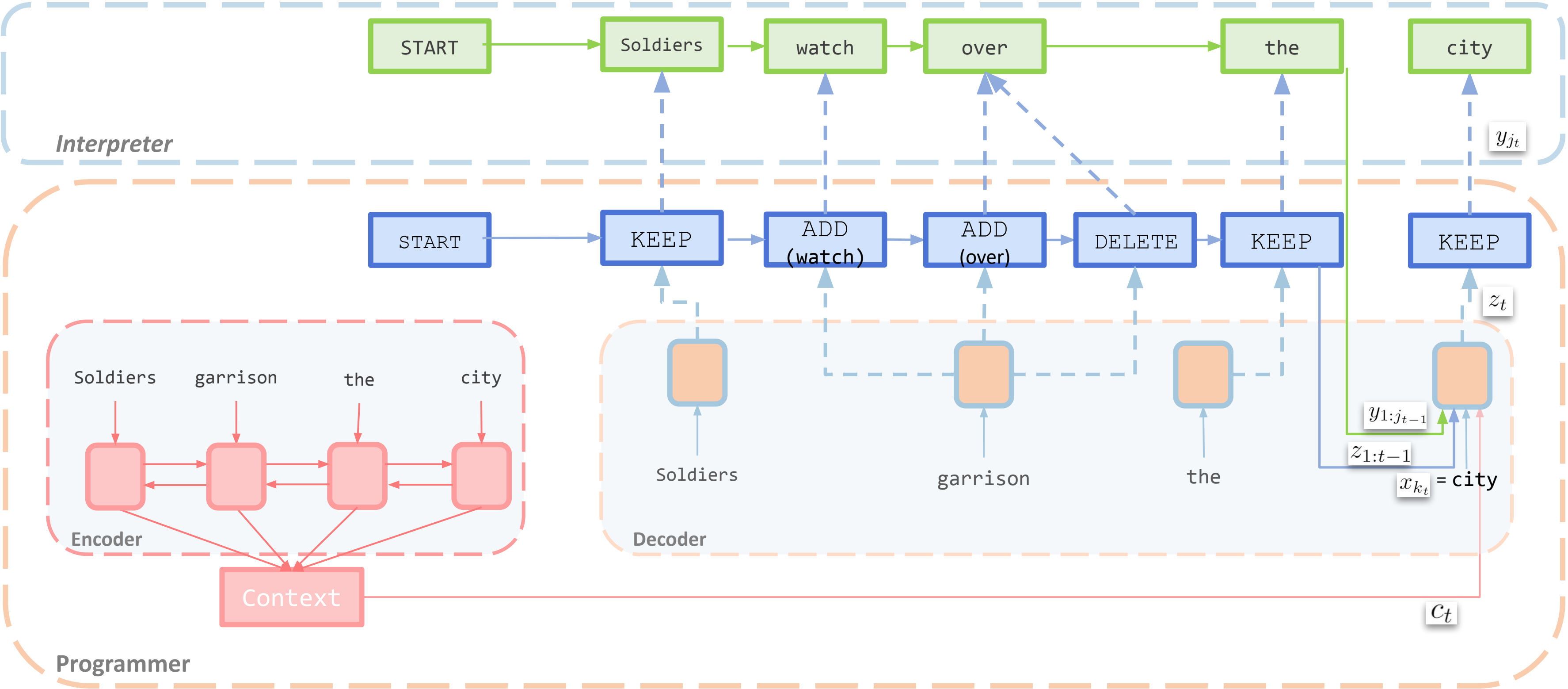
- Create edit labels explicitly:
  - through three types of edits (z): **ADD**, **DEL**, and **KEEP**

- New training objective function:
  - learn p(z|x)

Neural programmer-interpreter (NPI)

z

EditNTS: Walkthrough



Learning to transform input to output by **edit operations**.

Experiments & Results

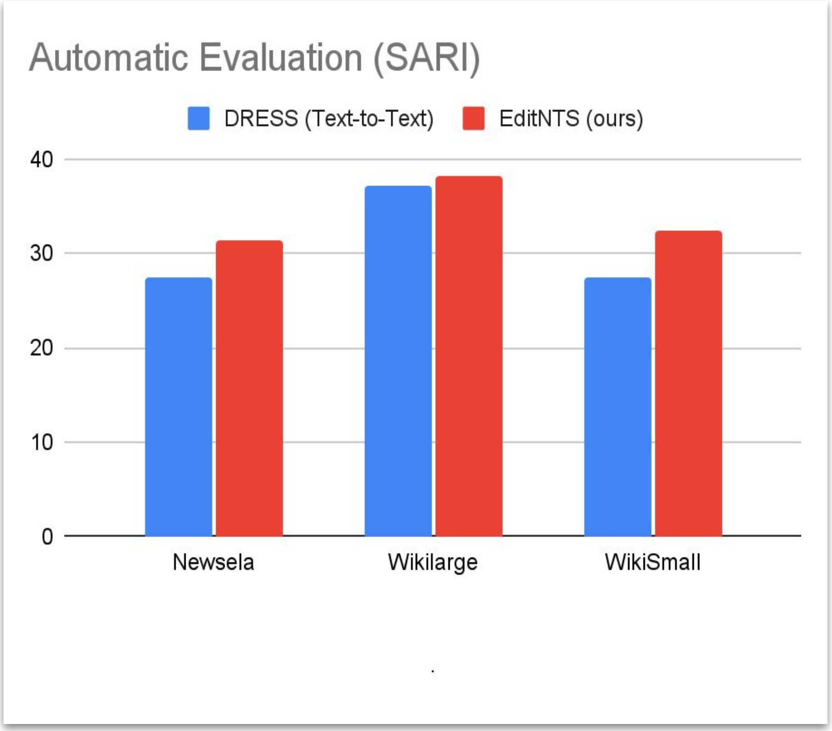
Compared to DRESS [1] (seq2seq) on Newsela, Wikilarge and Wikismall:

- SARI improvements by

  +4.04, +1.14, +4.87
- Prefered by human judges

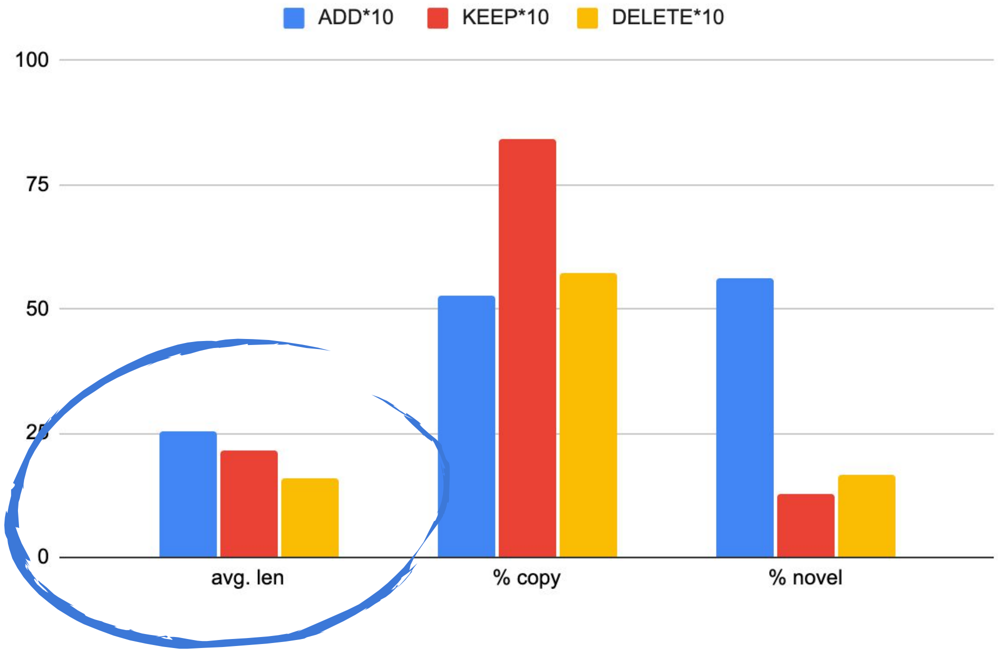
Facts and rare entities preserving by KEEP

[1] Zhang and Lapata. *Sentence Simplification with Deep Reinforcement Learning.* EMNLP 2018



SARI (Xu et al., 2016): measure similarity to both input and reference sentence

Controlled Generation with Edit Type Bias

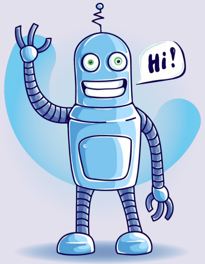


Reward ADD:
- **Long output**
- **More novel words**

Reward **KEEP**:
- More copy

Reward **DELETE**:
- Short output

Key Takeaways



- **Inductive bias of learning edits** can be useful for **faithful** and **controlled** generation
  - Important concepts can be directly kept
  - Output length, abstractive level, etc. can be controlled by associate costs with edit operations

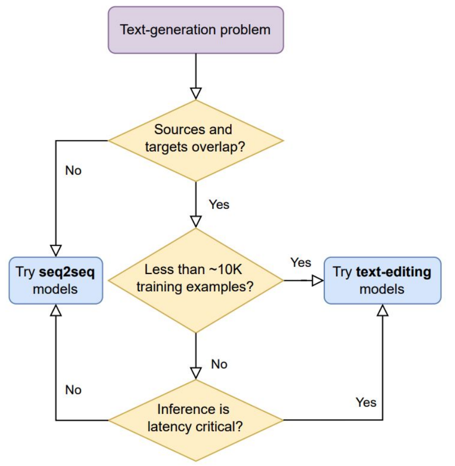

Text-generation problem
Sources and targets overlap?
No
Yes
Try **seq2seq** models
Less than ~10K training examples?
Yes
Try **text-editing** models
No
No
Inference is latency critical?
Yes

Figure 2: Proposed flowchart for deciding when to try a text-editing approach.

[1] Malmi, E., Dong, Y., Mallinson, J., Chuklin, A., Adamek, J., Mirylenka, D., ... & Severyn. *Text Generation with Text-Editing Models.* NAACL 2022 Tutorial

Faithful to the Document or to the World? Mitigating Hallucinations via Entity-Linked Knowledge in Abstractive Summarization

Yue Dong , John Wieting and Pat Verga

EMNLP 2022
Findings

Verify hallucination with world knowledge

Variants of hallucinations [1]

Intrinsic: generated text contradicts source text

vs.

Extrinsic: generated text is not grounded in the source text

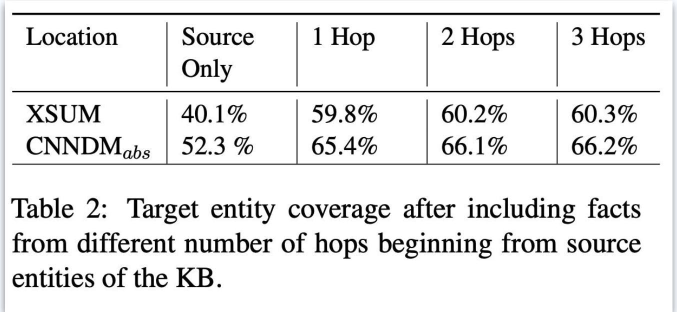
Factual: **extrinsic hallucination** consistent with world knowledge [2]

[1] Maynez et al., *On Faithfulness and Factuality in Abstractive Summarization*. ACL 2020.
[2] Cao et al., *Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization.* ACL 2022.

Human-Written Summaries Contain “Hallucination”

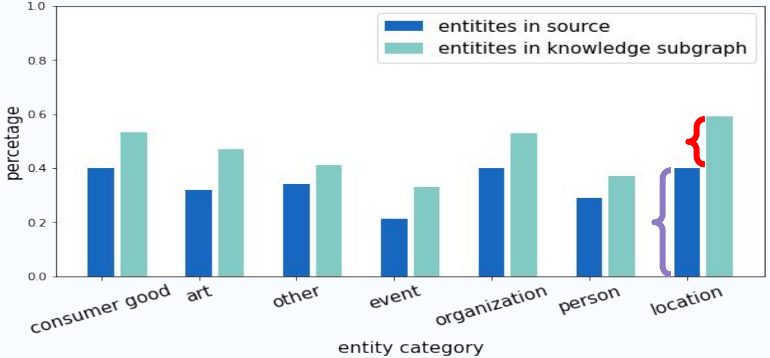
On Xsum and CNN_abs:

- **48%~60% of reference entities are not in the source**
- **Memorized (factual) knowledge** in humans
- **Many of them are one-hop facts!**
  Xsum, Location-based target entities:
  - 40% in the source
  - 20% in one-hop facts



Location	Source Only	1 Hop	2 Hops	3 Hops
XSUM	40.1%	59.8%	60.2%	60.3%
CNNDM$_{abs}$	52.3 %	65.4%	66.1%	66.2%

Table 2: Target entity coverage after including facts from different number of hops beginning from source entities of the KB.



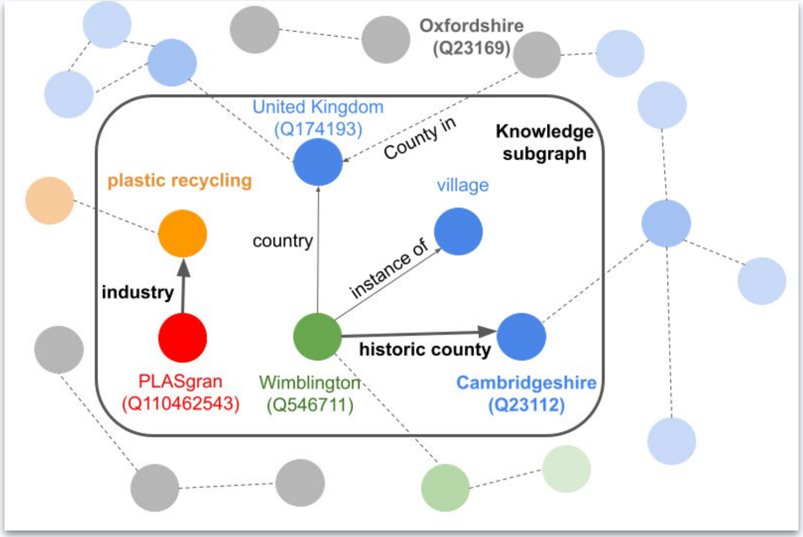
Constructing Knowledge Subgraph of A Document

Given a document,

1. Extracting all source entities

2. Including facts that are one-hop away on Wikidata

Document: A fire crew remains at Plasgran Wimblington. The incident began more than 16 hours ago. Road closures are expected …



Correct Factual Errors with World Knowledge

Input: A fire crew remains at Plasgran, Wimblington. The incident began more than 16 hours ago. Road closures are expected ...

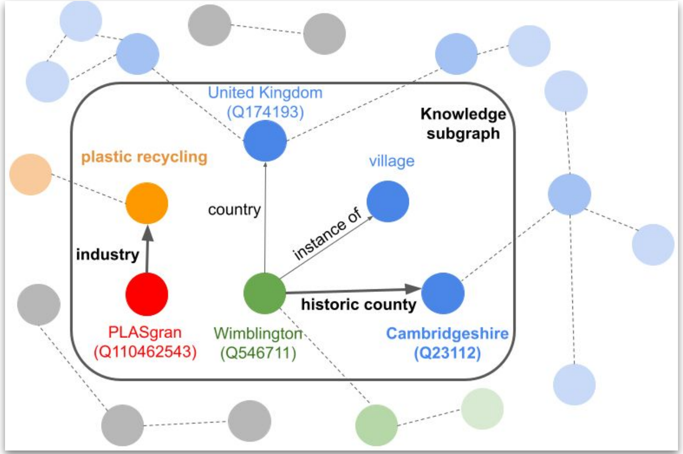
(A)

System-generated summary:
A large fire has broken out at a recycling centre in Oxfordshire...

Entity Masking **(B)**

Entity masking:
A large fire has broken out at a [MASK] in [MASK]...

(C) ***Facts Linking***



Knowledge Graph (**G**)

Entity Correction **(D)**

(E)

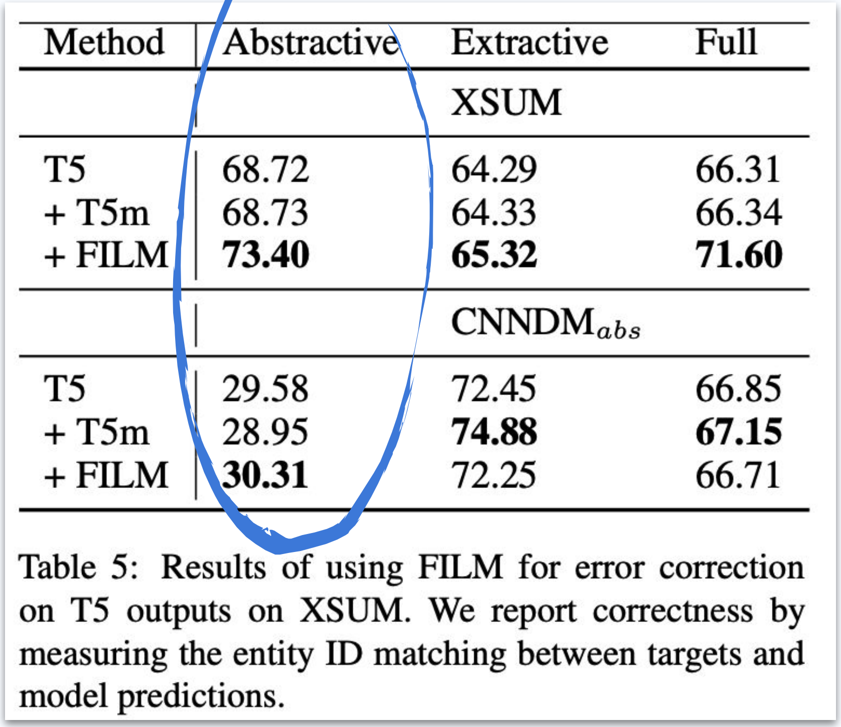
Memory

Summary with fact-based entity correction:
A large fire has broken out at a plastic recycling centre in Cambridgeshire...

Results and Factual Creativity

Using **one-hop facts**,

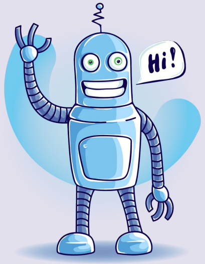
Models can generate more entities **matching human choices**



Method	Abstractive	Extractive	Full
		XSUM	
T5	68.72	64.29	66.31
+ T5m	68.73	64.33	66.34
+ FILM	**73.40**	**65.32**	**71.60**
		CNNDM$_{abs}$	
T5	29.58	72.45	66.85
+ T5m	28.95	**74.88**	**67.15**
+ FILM	**30.31**	72.25	66.71

Table 5: Results of using FILM for error correction on T5 outputs on XSUM. We report correctness by measuring the entity ID matching between targets and model predictions.

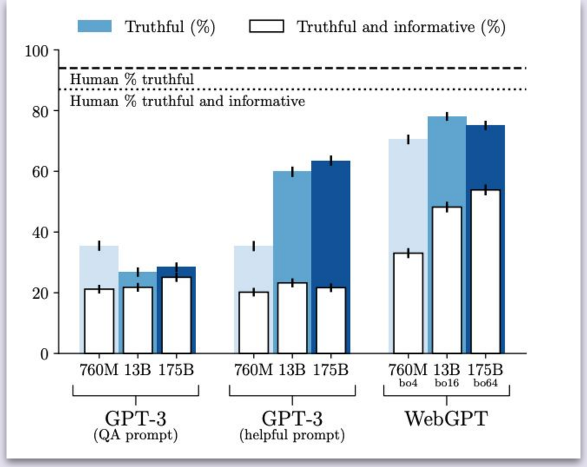
Key Takeaways



- **Not all hallucinations** are undesirable
  - Human written summaries contain **many one-hop extrinsic & factual hallucinations**
  - Suggest human using one-hop reasoning when summarizing articles?

- Inductive bias of **using symbolic knowledge base (KB)** allows models to generate more entities that **match human preferences**

Human imitation learning



[1] Nakano, Reiichiro, et al. "WebGPT: Browser-assisted question-answering with human feedback." OpenAI 2021

This Thesis

Designing models with **appropriate inductive bias** beyond the standard seq2seq

1. For Salience

Selecting **important information**

Seq2Set

2. For Faithfulness

consistent with **the source**

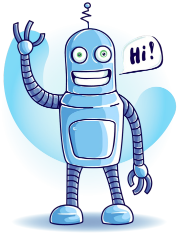
Seq2Edits

3. For Factuality

consistent with **the world knowledge**

Seq + Knowledge

Thank you!



Thanks to all my collaborators!

For a full list of my contributions, check out my website: https://yuedongcs.github.io/

@YueDongCS

✉ yue.dong@ucr.edu

Academic Collaborations:
Jackie Cheung, Meng Cao, Rui Meng, khushboo Thaker, Lei Zhang, Daqing He, Andrei Romascanu, Yao Lu, Laurent Charlin, Jiapeng Wu, Matt Grenander, Annie Louis, Pengfei Liu, Jie Fu, Xipeng Qiu, Yikang Shen, Eric Crawford, Herke van Hoof, Koustuv Sinha, Derek Ruths

Industrial Internships:
William Cohen, **Yejin Choi**, **Pat Verga, Chandra Bhagavatula, Jingjing Liu**, Shuohang Wang, Zhe Gan, Yu Cheng, John Wieting, Xingdi Yuan, Tong Wang, Ximing Lu, Jena D. Hwang, Antoine Bosselut, Zichao Li