Robust and Trustworthy NLP Through The Lens of Text Summarization

Yue Dong

McGill University x Mila

Natural language processing (NLP) offers incredible opportunities for automating tasks that involve human languages

QA

Chatbot

Translation

Summarization

Grammar Correction

Search

Conditional Text Generation

Generate text **y** according to some pre-specified conditioning **x**

Conditioning **x**:

- Machine Translation: source language
- Dialogue: previous turns
- Question Answering: question
- **Text Summarization:** source document

- **Fundamentally Challenging**
- **Practically Important**

Extractive summary:
A fire crew remains at Plasgran, Wimblington.

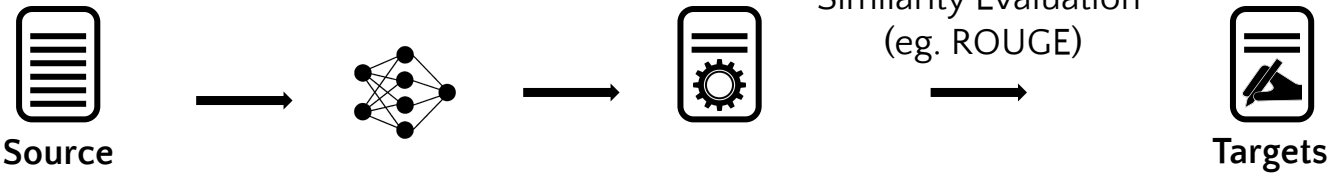
Source

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

Abstractive summary:
A large fire has broken out at Plasgran in Cambridgeshire.

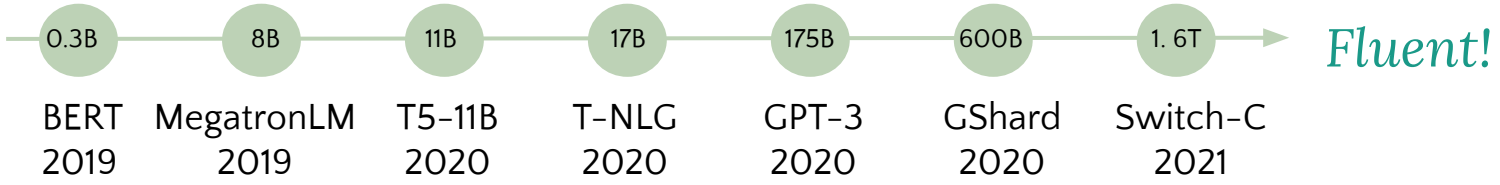
Summarization Requirements

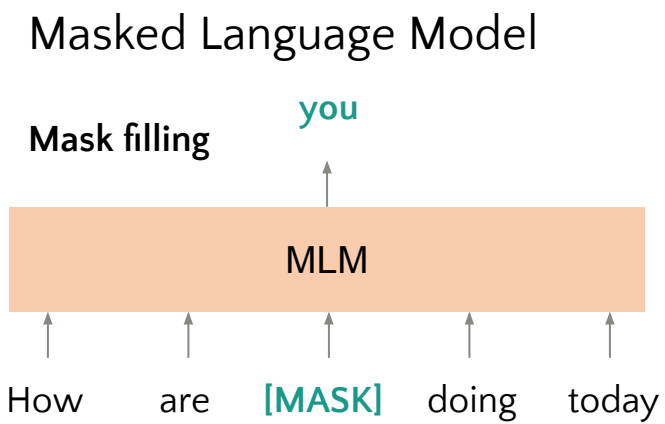
Shortening text while preserving main ideas

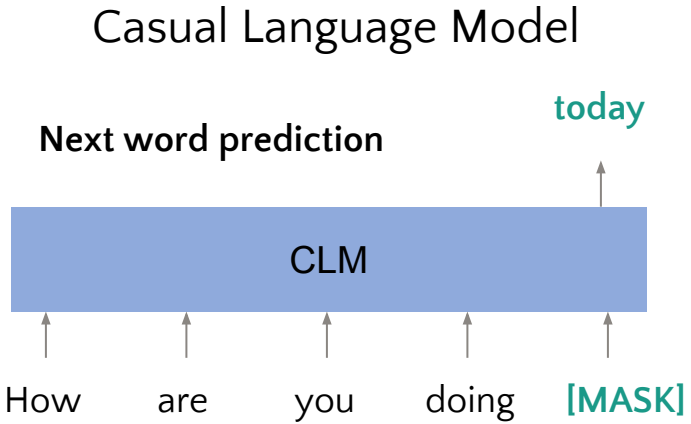


Pre-trained Language Models (PLMs)

Treating every text generation problem as a “text-to-text” problem.







Challenges: Limitations of Seq2seq

"The more technical the content … the greater the risk that the text the AI produces will contain flat-out wrong statements."

Review: Rytr across 20 NLP tasks

1. **Robust** to distribution shift
2. **Trustworthy** in generation

My interests: How to improve **robustness** and **trustworthiness** in NLP?

Robustness and Trustworthiness in NLP

Competitive approaches:

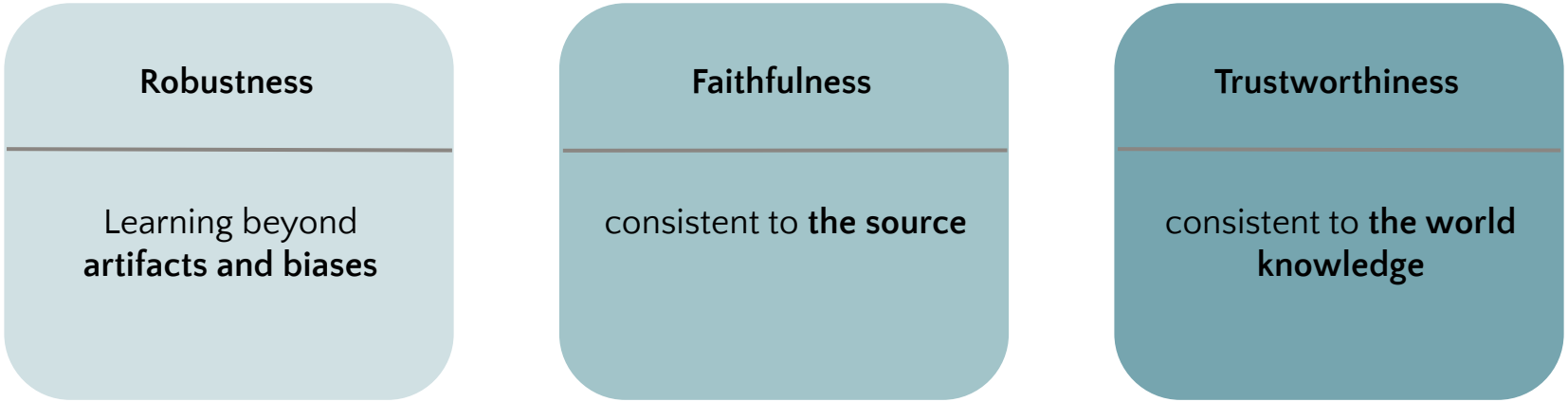
- Better PLMs with cleaner data and bigger model.
- Converting more NLP tasks to seq2seq.

Q: Everything can be learned by **seq2seq?**

My proposal:

Setups beyond **standard Seq2seq** are also important!

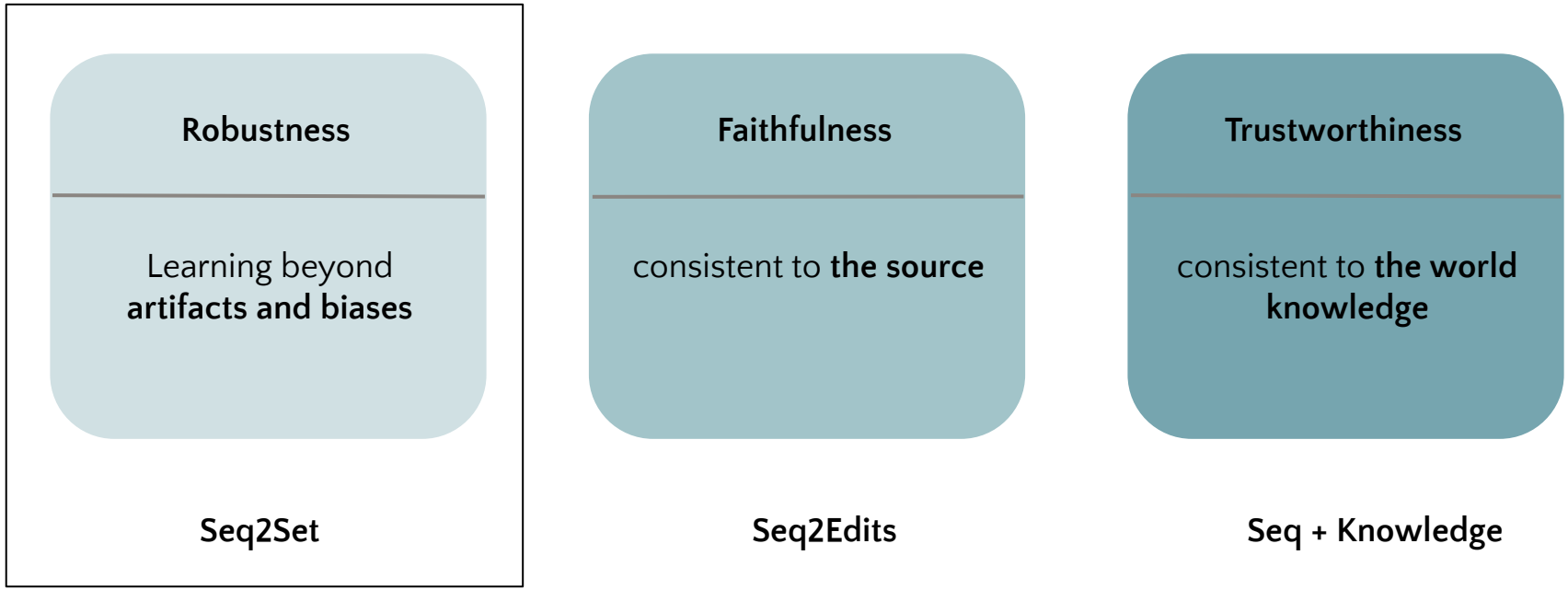
Talk Outline



Seq2Set **Seq2Edits** **Seq + Knowledge**

Setups beyond **standard Seq2seq** are also important!

Part I



Setups beyond **standard Seq2seq** are also important!

Artifacts & Biases

Source

Similarity Evaluation
(eg. ROUGE)

Targets

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

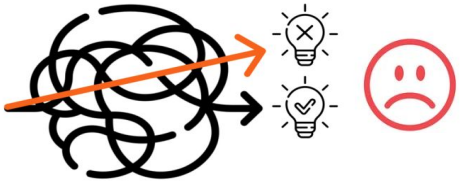
A fire crew remains at **Plasgran**, **Wimblington**.

*A fire has damaged **a plastics factory** in **Cambridgeshire**.*

Chooses the 1st sentence as a summary

Artifacts & Biases

Always picks the 1st sentence



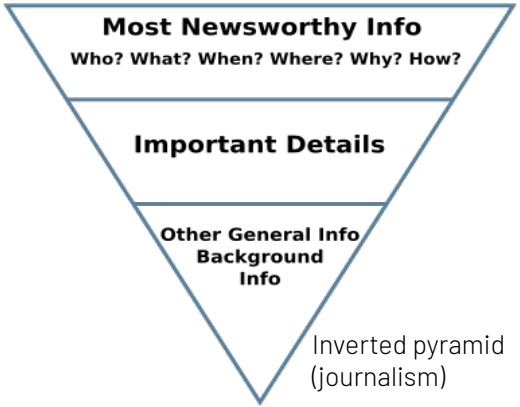
Contents

1st sentence is important in this example

What Biases?

Domain-specific text structures biases often give shortcuts for **picking important information**.

News

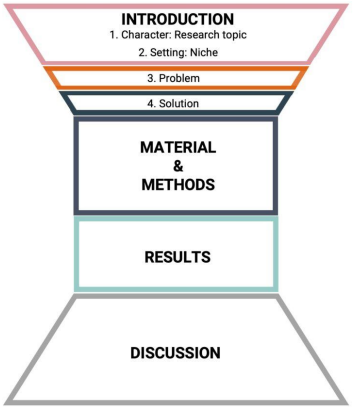


[Dong et al., EMNLP 18, EMNLP 19]

Stories

[Dong et al., ACL 21]

Scientific Articles

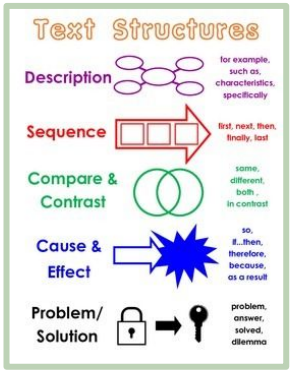


[Dong et al., EACL 21]
[*, Dong et al., EMNLP 20, ACL 21]

Lead Bias in Extractive News Summarization

Lead: Bangladesh beat fellow World Cup quarter-finalists Pakistan by 79 runs in the first one-day international in Dhaka. Rahim scored centuries as Bangladesh made 329 for six and Pakistan could only muster 250 in reply.

Reference: Bangladesh beat fellow World Cup quarter-finalists Pakistan by 79 runs. Rahim scored centuries for Bangladesh. Bangladesh made 329 for six and Pakistan could only muster 250 in reply.

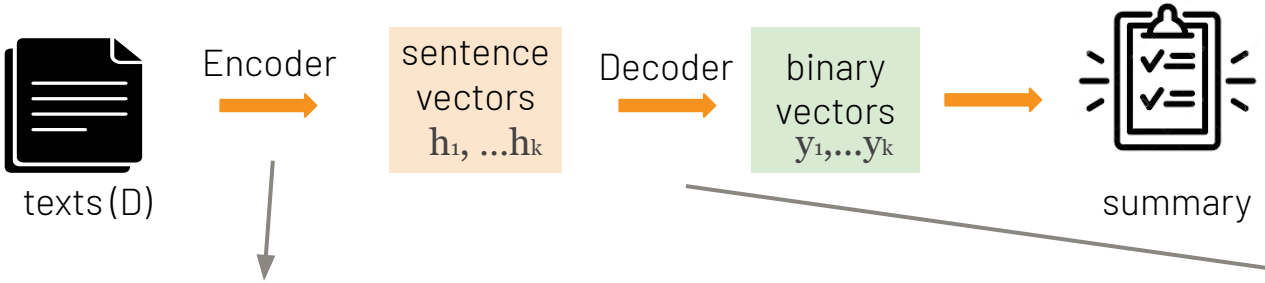


Lead: Standing up for what you believe. What does it cost you? What do you gain?

Reference: Indiana town's Memories Pizza is shut down after online threat. Its owners say they'd refuse to cater a same-sex couple's wedding.

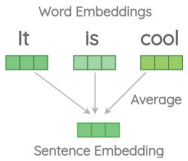
More than **20-30%** of summary-worthy sentences come from the second half of news documents (Nallapati et al., 2017; Kedzie et al., 2018)

Autoregressive Sequential Tagging Models



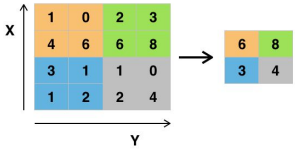
Encoder

Pooling



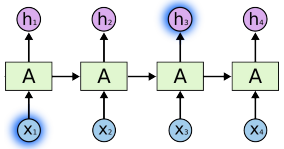
[Kageback et al.,2014]

CNNs



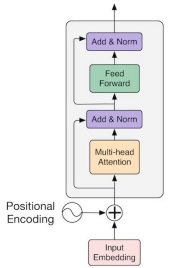
[Yin and Pei, 2015, [Cao et al., 2015]

RNNs



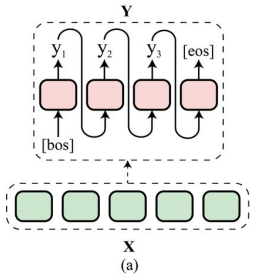
[Cheng and Lapata, 2016, Nallapati et al., 2017, Zhou et al., 2018]

Transformers



[Zhang et al., 2019, Liu and Lapata, 2019, Zhong et al., 2020]

Autoregressive decoder

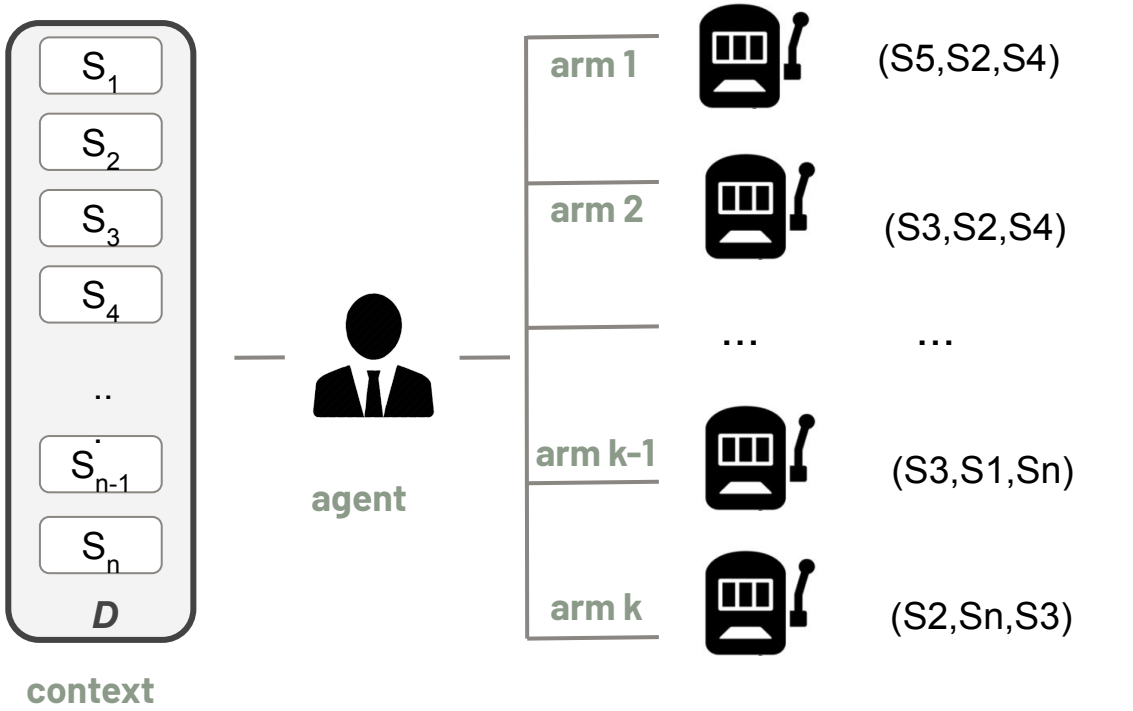


$$y_t = f(D, h_t, y_{1:t-1})$$

Is extractive sequential tagging **robust in distribution shifts**?

BanditSum: Break into Non-autoregressive

Trained by REINFORCE to directly optimize **content importance**, regardless of position in the document



Contextual Multi-armed Bandit

Context = the document

Arm = a set of *M* sentences

Reward = f (arm, context)

Selection is not sequential,
Learns appropriate lead bias

Dong, Shen, et al., BanditSum: Extractive Summarization as a Contextual Bandit. EMNLP (2018)

BanditSum: RL in a Nutshell

Goal: maximize reward R
system summary **i**, reference summary **a**

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

In BanditSum for summarization:

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

Optimization problem:
computation of ROUGE is not differentiable!

How? policy gradient reinforcement learning
likelihood ratio gradient estimator (Williams, 1992)

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

Sample batch size = B for expectation estimation:

$$\nabla_\theta J(\theta) \approx \frac{1}{B}\sum_{b=1}^{B} \nabla_\theta \log p_\theta(i^b|d) R(i^b,a) \qquad (3)$$

Policy needs to be differentiable!

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

Deterministic $\pi_\theta(d)$

texts (D)

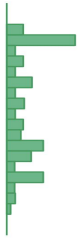
Encoder

content importance

vectors

$\pi_1, \ldots \pi_k$

$\pi_i \in [0,1]$



Context - State

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

~~Decoder~~

Sampling

summary

Arm - Action

Sampling without replacement

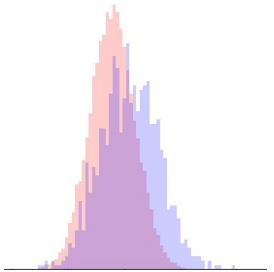
$$\prod_{j=1}^{M} \left(\frac{\epsilon}{N_d - j + 1} + \frac{(1-\epsilon)\pi(d)_{i_j}}{z(d) - \sum_{k=1}^{j-1} \pi(d)_{i_k}} \right)$$

ε-greedy sampling

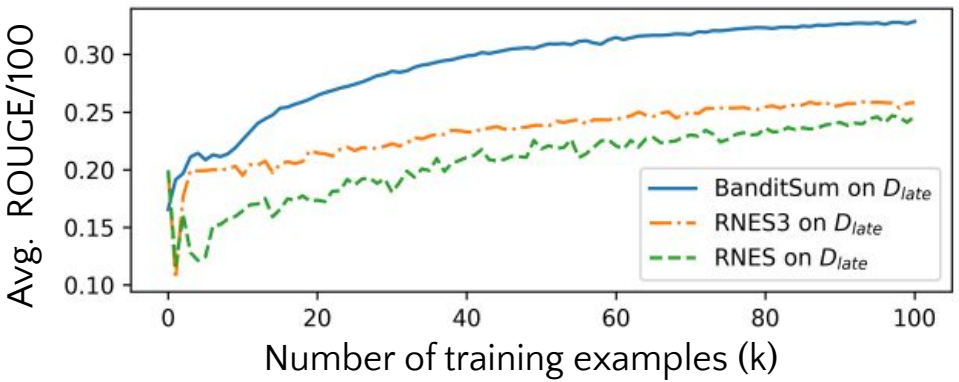
Robustness in Extractive Summarization

Dataset: CNN/DailyMail, 287k/13k/11k document-summary pairs
ROUGE: similarity between generated summary and gold-reference summary
Domain shift in Test (3.8k)**:**

- With different lead bias distribution
- D_{late} : summary-worthy sentences appear **late** in the article

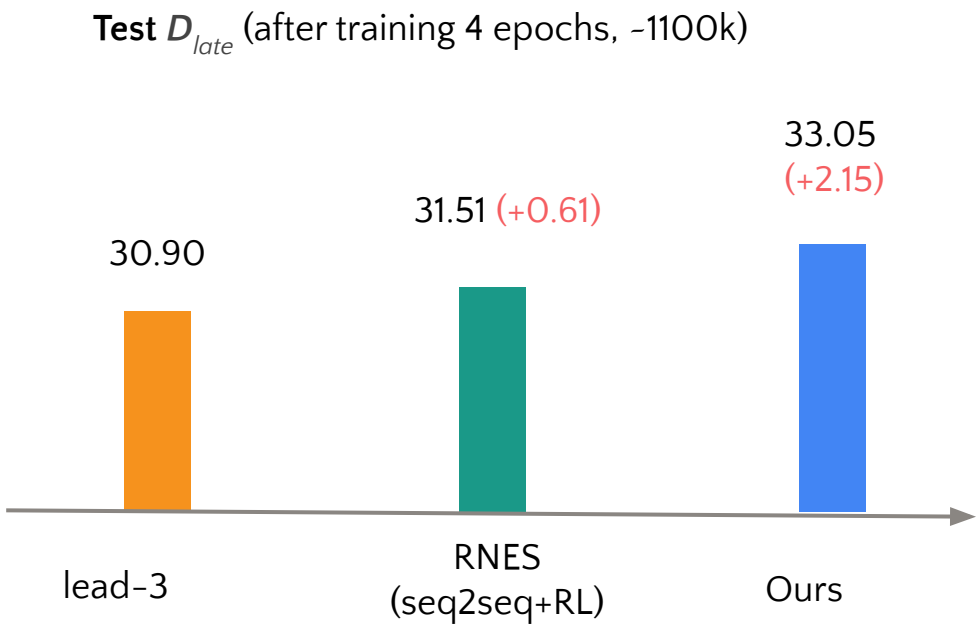


Validation D_{late}



Efficient Learning!

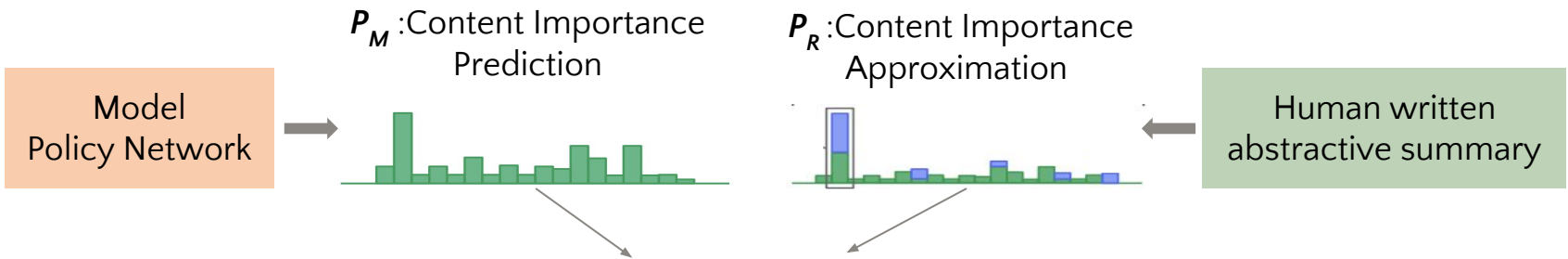
Robustness in Extractive Summarization



Better **overall ROUGE performance** and **human ratings** on the full test set.

Inductive Prior for Content Importance

Adds "**content importance**"-based **entropy regularization** in RL training



$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

Encourage model predictions P_M to match P_R :

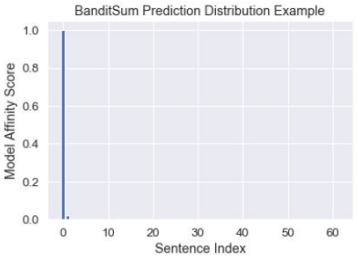
$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(P_R \parallel P_{\mathcal{M}})$$

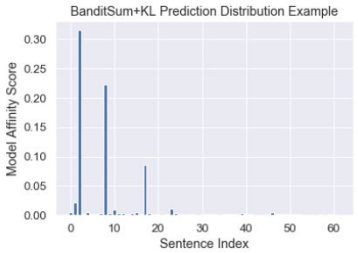
Dong, Grenander, et al., Countering the Effects of Lead Bias in News Summarization via Multi-Stage Training and Auxiliary Losses. EMNLP (2019)

Inductive Prior as Regularization

Use auxiliary loss (regularization) when training the model:

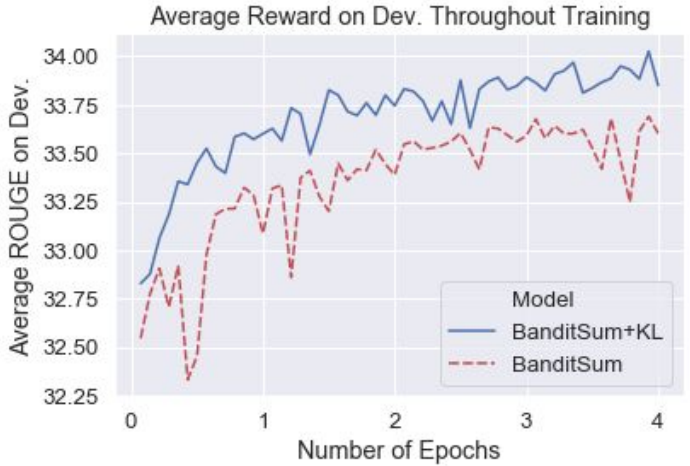
$$\theta^{(t+1)} = \theta^{(t)} + \alpha \left(\nabla \mathcal{L}_{\mathcal{M}}(\theta^{(t)}) + \beta \nabla \mathcal{L}_{\mathrm{KL}}(\theta^{(t)}) \right)$$





Low entropy

High entropy



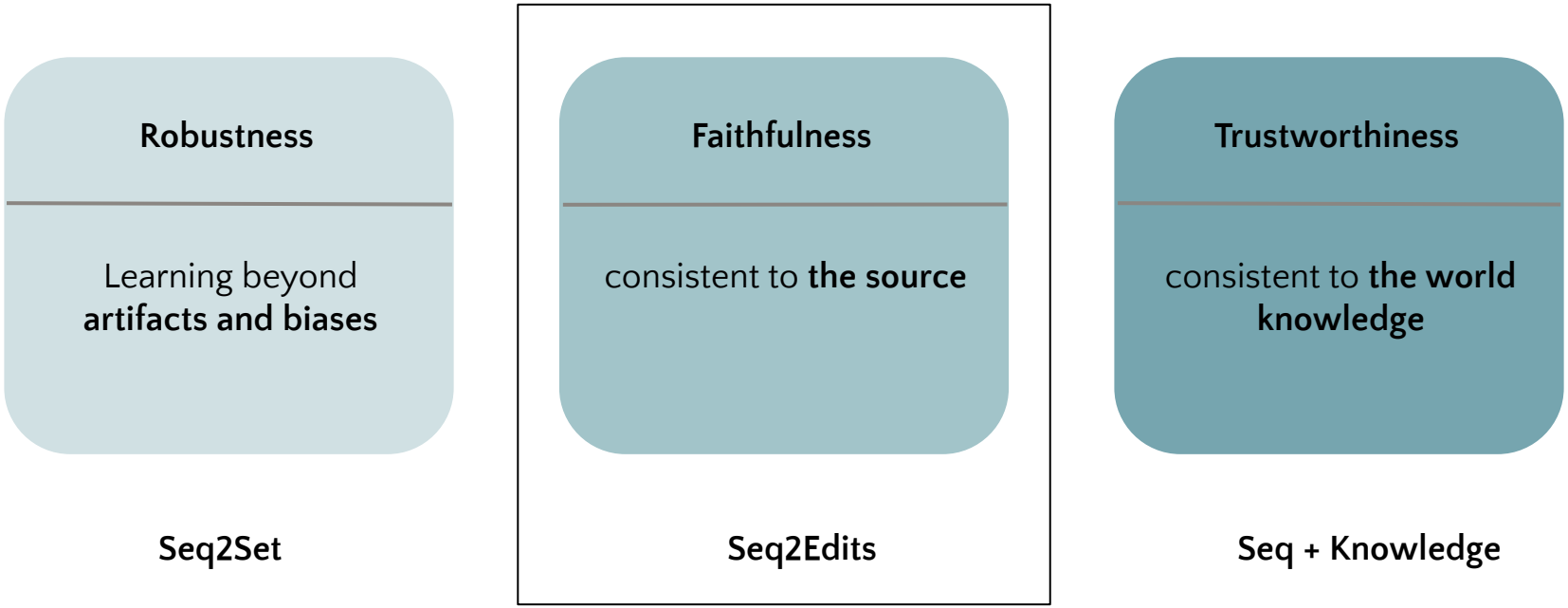
Better Content Importance Learning in RL

Part I: Robustness

Takeaways:

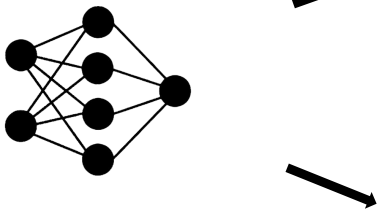
Setups beyond **standard Seq2seq** are important for learning beyond artifacts and biases (seq2sets)

Part II



Setups beyond **standard Seq2seq** are also important!

Comparing with Targets is Not Enough



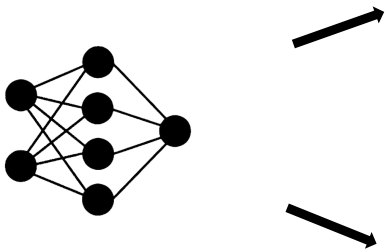
A fire crew remains at Plasgran, Wimblington.

A fire crew remains at Plasgran, Oxfordshire.

Targets

A fire has damaged a plastics factory in Cambridgeshire.

Faithful to The Source



A fire crew remains at Plasgran, Wimblington.

A fire crew remains at Plasgran, Oxfordshire.

Source

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

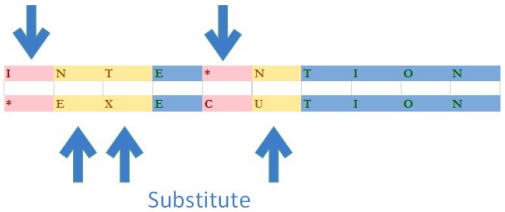
Hallucination: Generation that is not backed up by the source.

Why Models Hallucinate?

Seq2seq models are prone to hallucination due to a large generation freedom (Xiao et al. 2021).

Our proposal (Seq2Edits):

Bounds the generation freedom with edit-distance by learning edits



Dong el al., EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. ACL (2019)

EditNTS: Edit-based Training

- 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)

$$P(\mathbf{z}|\mathbf{x}) = \prod_{t=1}^{\mathbf{z}} P(z_t|y_{1:j_{t-1}}, z_{1:t-1}, x_{k_t}, \mathbf{x})$$

Dong el al., EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. ACL (2019)

Experiments

Datasets
(supervised, document - summary pairs)

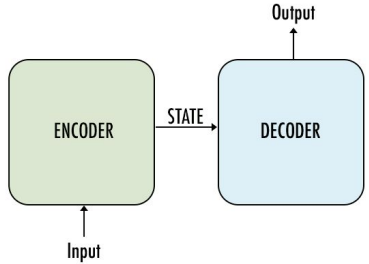
Models

WikiLarge & Small

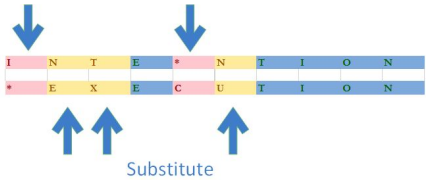
296,402/2000/359
& 88,837/205/100

newsela

94,208/1129/1076



Baselines: DRESS
(Zhang and Lapata, 2018)
best seq2seq models

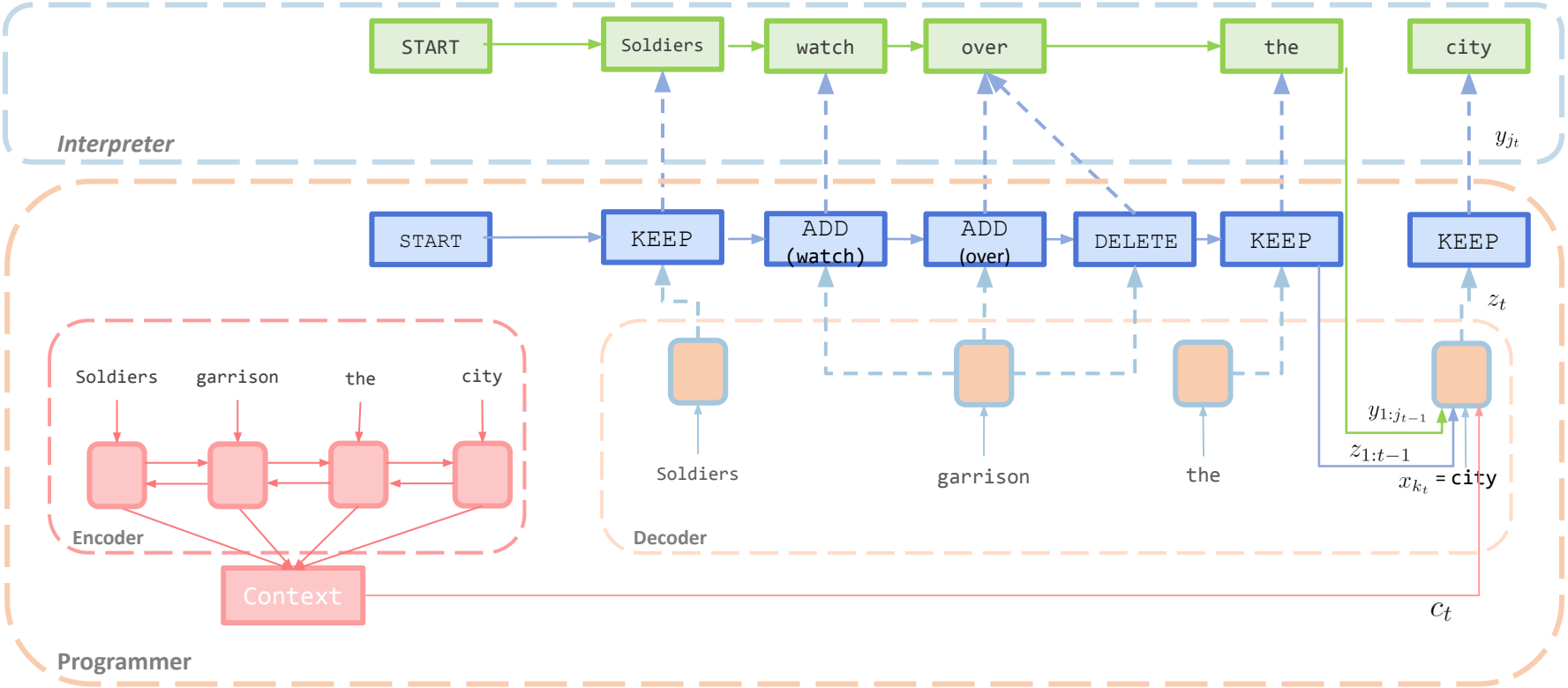


Ours: EditNTS
seq2edits with NPI

Evaluation:

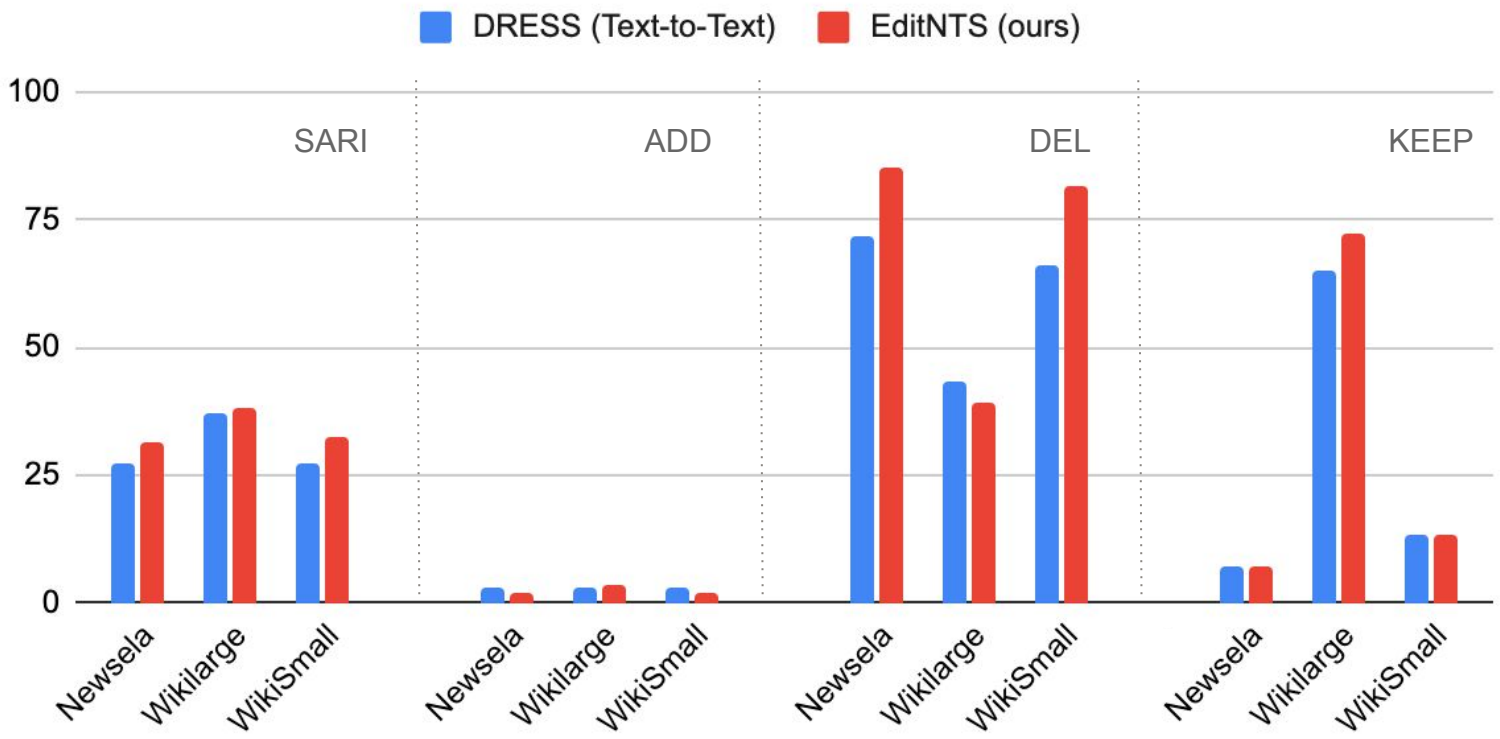
- SARI (Xu et al., 2016): Measure similarity to both input and reference sentence
- Three human judges rate based on fluency, adequacy, simplicity (a five-point Likert scale)

EditNTS: Walkthrough



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

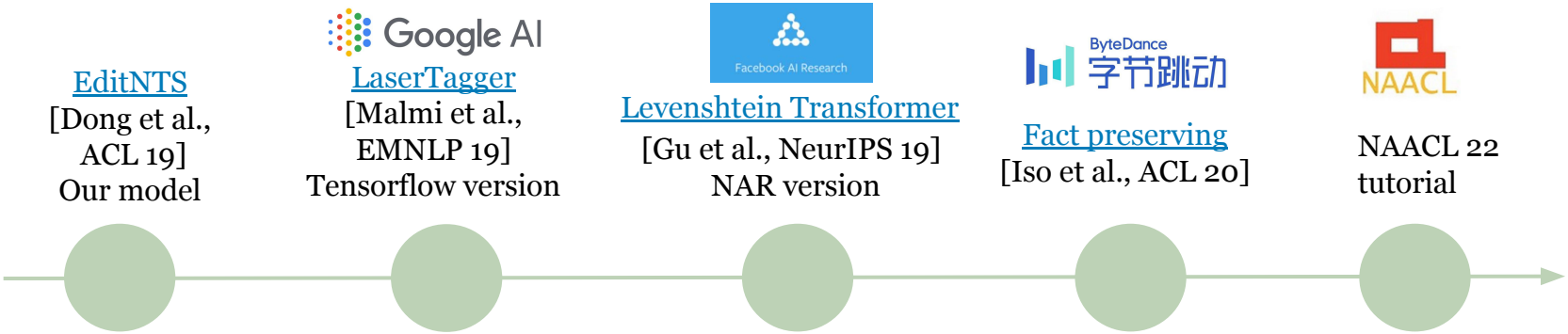
Automatic Evaluations



Benefits #1: Fact preserving by KEEP

Benefits #2: Controlled text generation by edit cost

Edit-Based Text Generation Models



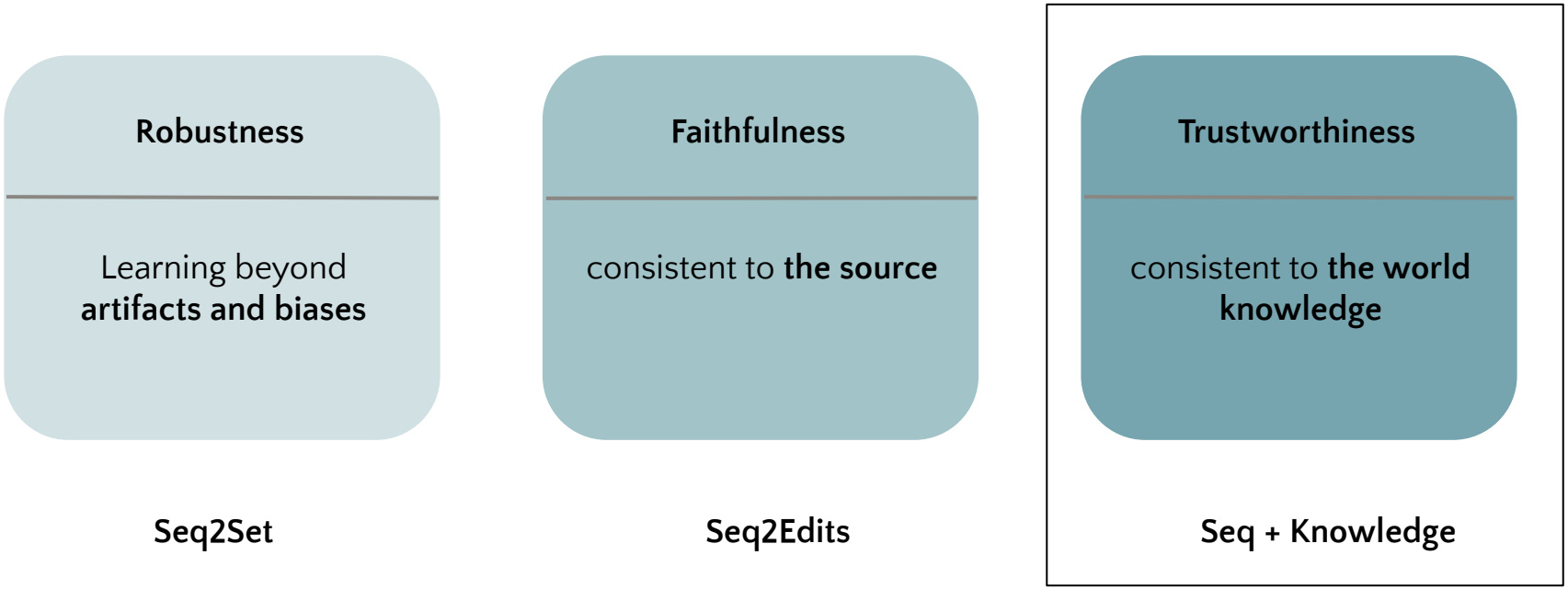
Benefits #1: Edit-based models increase FIDELITY by 14% F1 scores on different datasets
Benefits #2: Better human ratings in many tasks with large input/output overlap

Part II: Faithfulness

Takeaways:

Setups beyond **standard Seq2seq** are important for **bounding the generation freedom (seq2edits)**

Part III



Setups beyond **standard Seq2seq** are also important!

Are Hallucinations All Bad?

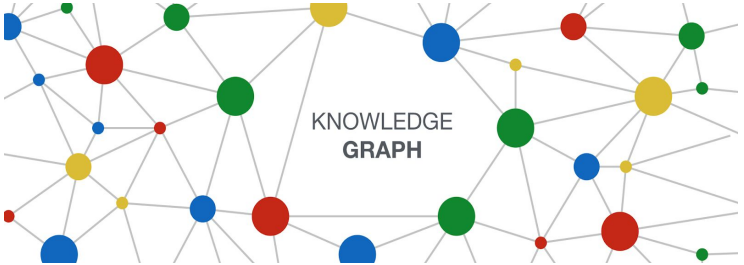
Hallucinations: generations that are not backed up by the source.

Input: In 2005 , A fire crew remains at Plasgran, Wimblington. The incident began more than 16 hours ago. Road closures are expected ...

Seq2seq model: A large fire has broken out at a recycling centre in Oxfordshire.

Reduce hallucinations directly as in part II (bounds the generation).

However, how do we know the relation between Oxfordshire and Wimblington?



External Knowledge!

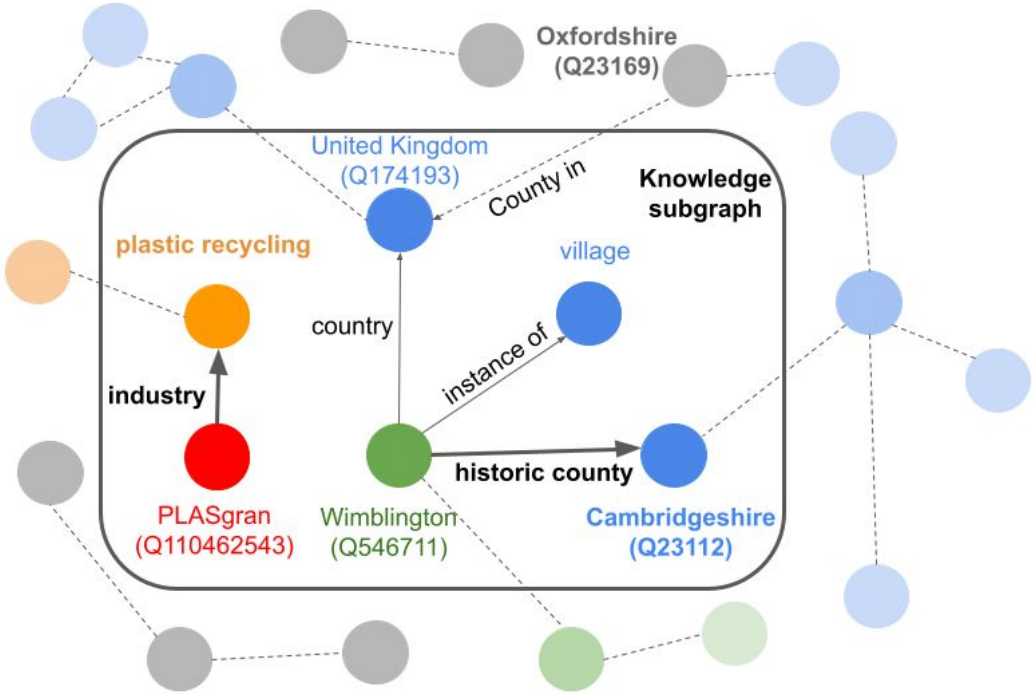
Constructing Knowledge Subgraph of A Document

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

Constructing knowledge subgraph

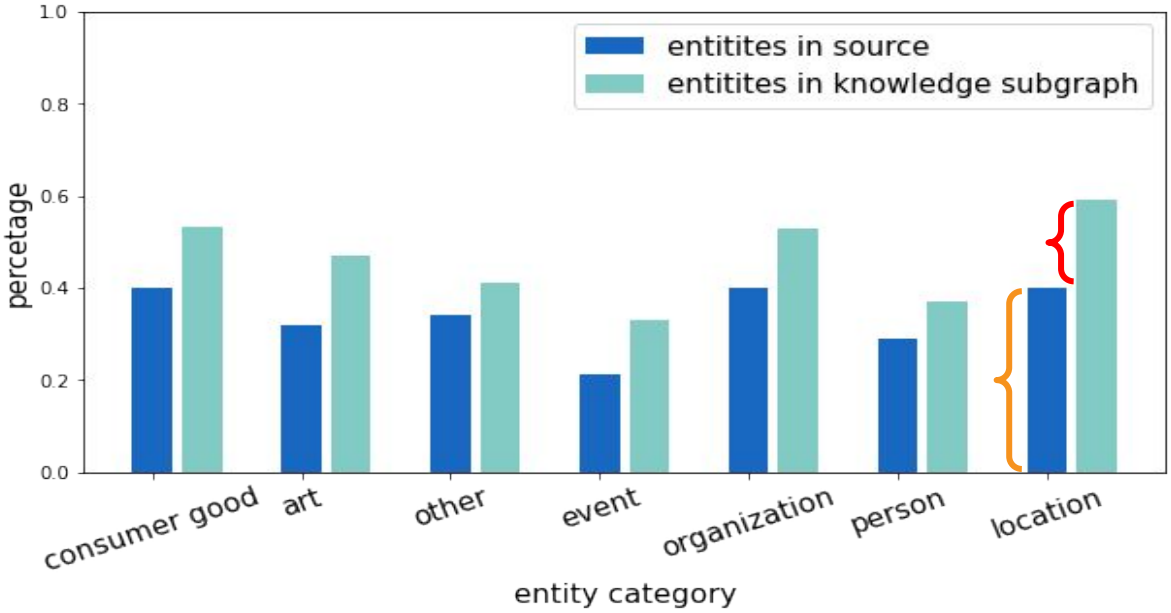
- Extracting all source entities
- Including facts that are one-hop away

Gold-reference summary written by human:
A fire has damaged a plastics factory in Cambridgeshire.



[Dong et al., in submission 22]

Many Hallucinations Are Backed by Knowledge



E.g.
Location-based target entities in the dev. documents:

- 40% in the source
- 20% in the knowledge subgraph
- 40% neither

In XSUM, many target entities are not in the source, but in the knowledge subgraph.

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)

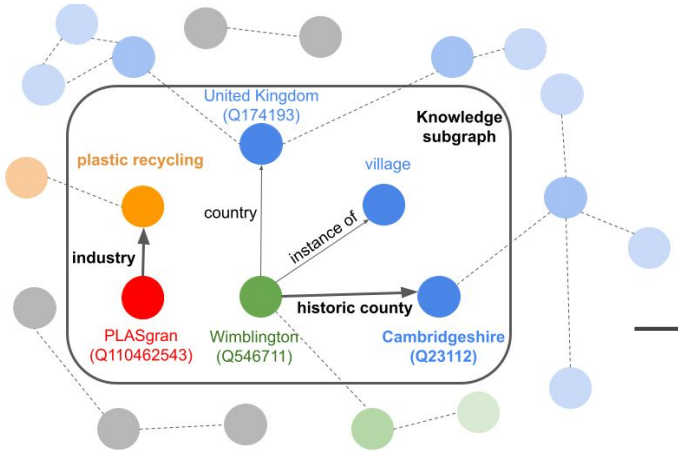
System-generated summary:
A large fire has broken out at a recycling centre in Oxfordshire...

Entity Masking (B)

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

Entity Correction (D)

(C) *Facts Linking*



Knowledge Graph (**G**)

(E)

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

Part III: Trustworthiness

Takeaways:

Setups beyond **standard Seq2seq** are important for **incorporating external knowledge (seq + knowledge)**

Summary: Setups Beyond Standard Seq2seq Is Also Important!

Robustness	Faithfulness	Trustworthiness
Learning beyond **artifacts and biases**	Generation is **consistent to the source**	Generation is **consistent to the world knowledge**
Seq2Set	**Seq2Edits**	**Seq + Knowledge**

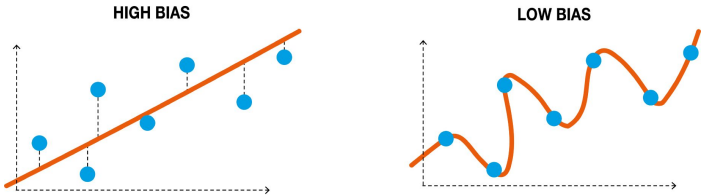
Benefit #1: Learning appropriate structure biases
Benefit #2: Bounds the generation freedom by edit distances and knowledge

Other Past Work

Robustness

Scientific & Medical Documents Summarization

1. Multi-documents [Lu, Dong et al., EMNLP 20]
2. Medical journals [*, Dong et al., ACL 21]



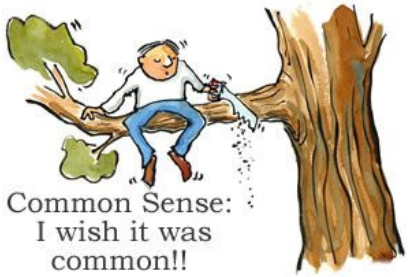
Inductive Bias Learning for Summarization

3. Entropy [Dong et al., EMNLP 19]
4. Discourse [Dong et al., EACL 21]

Trustworthiness

Faithful Summarization

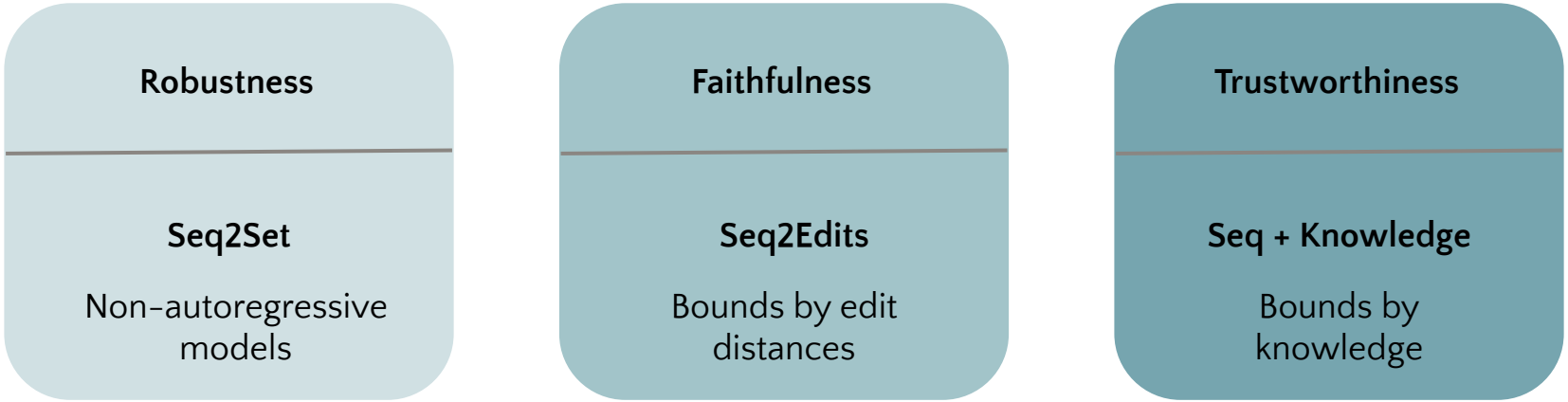
5. QA [Dong et al., EMNLP 20]
6. Adversarial data [Cao, Dong, et al., EMNLP 20]
7. Facts Prior [Cao, Dong, et al., ACL22]



Commonsense Reasoning for Generation

8. Attention patterns [Dong et al., ACL 21]

Short-Term Future Research: Beyond Standard Seq2seq



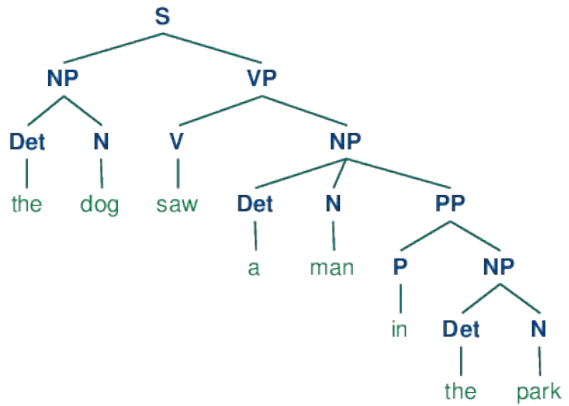
RL, Loss & Optimization, Inductive priors, Multimodality, Adversarial attacks

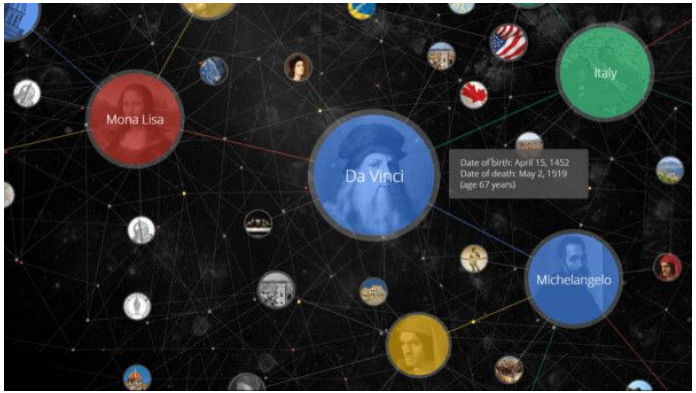
Uncertainty, Scaling, Theoretical bounds

Knowledge-enhanced generation, **Memory,** Information retrieval, Knowledge representation

Long-Term Future Research: Language

Q: Everything can be learned by **seq2seq?**



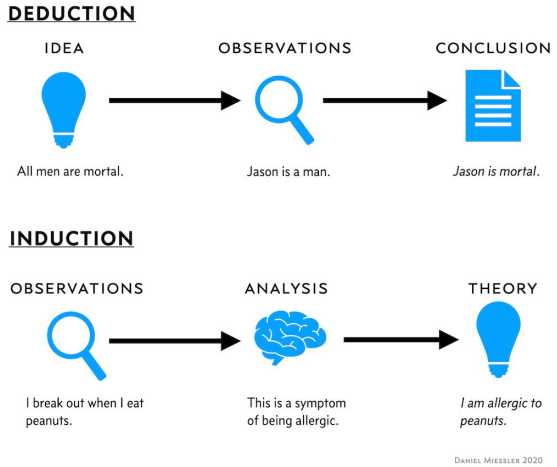


Linguistic forms
Morphology, syntax, semantics

Understanding
Local & world knowledge, common sense

Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *ACM FAccT 2021*

Even Longer-Term Future Research: Language & Intelligence



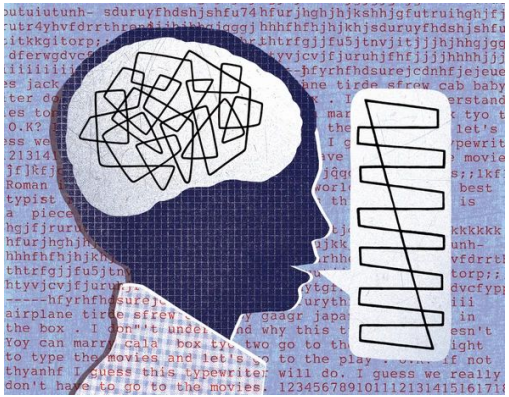
Reasoning
Causal relation, logics

And Beyond
Moral, emotional, ethics, cultural, fairness

Q: Can these all be learned by **seq2seq setup** **(next word prediction & mask filling)** with larger models and bigger & cleaner data?

Neuroscience Perspectives

Language is data emitted from the brain



To share thoughts with conspecifics

Language is for **effective communication**

To develop more complex thoughts

Language is not suitable for **complex thoughts**

Credit: Evelina Fedorenko

Future Research: Language & Intelligence

Language is for **effective communication**

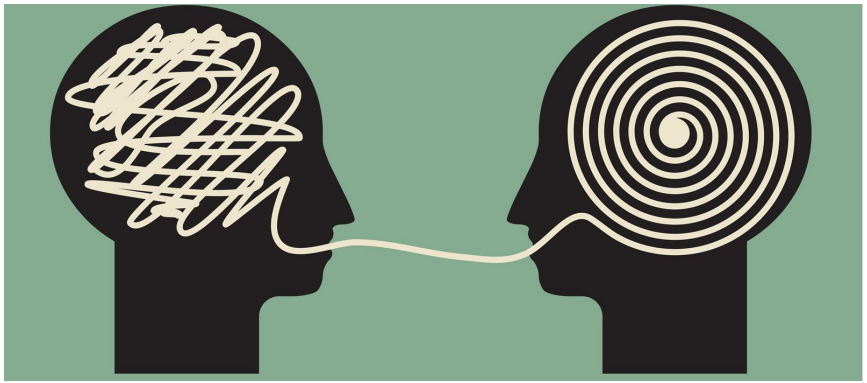
Language is not suitable for **complex thoughts**

Language modeling
(**Seq2seq:** next word prediction)

combine

Reasoning modeling
(**Structure prediction** beyond Seq2seq)

Seq2seq
has large
NLP capacity



Setups beyond
standard Seq2seq
are also important

Other Interests: Interdisciplinary Applications

NLP for Healthcare

NLP for Finance

NLP for Security

NLP for Education

Multimodal Learning

Fairness & Social Good

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

THANKS!

Industrial Internships:

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

Workshop & Tutorial Co-organizers:

1. Efficient Natural Language and Speech Processing (NeurIPS 21)

Mehdi Rezaghoizadeh, Lili Mou, Pascal Poupart, Ali Ghodsi, Qun Liu

2. New Frontiers in Summarization (EMNLP 21)

Wang Lu, Fei Liu, Jackie Cheung, Giuseppe Carenini

3. Text Generation with Text-Editing Models (NAACL 22)

Eric Malmi, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Daniil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, Aliaksei Severyn