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.
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

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

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

Abstractive summary:
A large fire has broken out at Plasgran in Cambridgeshire.
A fire crew remains at Plasgran, Wimblington. The incident began more than 16 hours ago. Road closures are expected...

1. Important based on targets
2. Faithful to the source
3. Trustworthy by world knowledge

E.g. Wimblington -> village in -> Cambridgeshire -> county in -> UK
Pre-trained Language Models (PLMs)

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

**Masked Language Model**
- Mask filling
  - How are you doing today

**Casual Language Model**
- Next word prediction
  - How are you doing [MASK] today
Challenges: Limitations of Seq2seq

“\textbf{The more technical the content} ... \textbf{the greater the risk} that the text the AI produces will contain flat-out \textbf{wrong} statements.”

Review: Rytr across 20 NLP tasks

1. \textbf{Robust} to distribution shift
2. \textbf{Trustworthy} in generation

My interests: How to improve \textbf{robustness} and \textbf{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

- **Robustness**: Learning beyond artifacts and biases
- **Faithfulness**: Consistent to the source
- **Trustworthiness**: Consistent to the world knowledge

Seq2Set  
Seq2Edits  
Seq + Knowledge

Setups beyond standard **Seq2seq** are also important!
Setups beyond standard Seq2seq are also important!
A fire crew remains at Plasgran, Wimblington. The incident began more than 16 hours ago. Road closures are expected …

A fire has damaged a plastics factory in Cambridgeshire.

Choose the 1st sentence as a summary

Always picks the 1st sentence

1st sentence is important in this example
What Biases?

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

News

Most Newsworthy Info

Important Details

Other General Info
Background Info

Inverted pyramid (journalism)

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

Stories

RISING ACTION
We read about the main events leading up to the problem.

BACKGROUND KNOWLEDGE
We meet and get to know the characters and learn about the setting.

CLIMAX
We find out the main problem of the story. This is when the biggest action happens and changes the course of the story.

FALLING ACTION
The problem begins to get solved.

SOLUTION & CONCLUSION
The problem is solved.

[ Dong et al., ACL 21 ]

Scientific Articles

INTRODUCTION
1. Character: Research topic
2. Setting: Niche

3. Problem

4. Solution

MATERIAL & METHODS

RESULTS

DISCUSSION

[ 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

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.

BanditSum: RL in a Nutshell

Goal: maximize reward $R$

system summary $i$, reference summary $a$

$$J(\theta) = E[R(i, a)]$$  \hspace{1cm} (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)

$$\nabla_\theta J(\theta) = E[\nabla_\theta \log p_\theta(i|d)R(i, a)]$$  \hspace{1cm} (2)

Sample batch size $= B$ for expectation estimation:

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

Policy needs to be differentiable!
Structure of Policy $p_\theta(\cdot|d) = \mu(\cdot|\pi_\theta(d))$

Deterministic $\pi_\theta(d)$

Encoder

texts (D)

Context - State

$\tau_1, \ldots \tau_k$

$v_\pi \in [0,1]$

Sampling without replacement

\[
\prod_{j=1}^{M} \left( \frac{\epsilon}{N_d - j + 1} + \frac{(1-\epsilon)\pi(d)_{ij}}{z(d) - \sum_{k=1}^{j-1} \pi(d)_{ik}} \right)
\]

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

-Decoder

Sampling

summary

Arm - Action

Sampling with $\epsilon$-greedy sampling

$\epsilon$-exploitation

$\epsilon$-exploration
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

**Efficient Learning!**

![Efficient Learning Image]
Robustness in Extractive Summarization

Test $D_{late}$ (after training 4 epochs, ~1100k)

- **lead-3**: 30.90
- **RNES (seq2seq+RL)**: 31.51 (+0.61)
- **Ours**: 33.05 (+2.15)

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

$P_M$ : Content Importance Prediction

$P_R$ : Content Importance Approximation

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

Encourage model predictions $P_M$ to match $P_R$:

$\mathcal{L}_{KL} = D_{KL}(P_R \parallel P_M)$

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:

\[ \theta(t+1) = \theta(t) + \alpha \left( \nabla \mathcal{L}_M(\theta(t)) + \beta \nabla \mathcal{L}_{KL}(\theta(t)) \right) \]

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

Robustness
Learning beyond artifacts and biases

Faithfulness
consistent to the source

Trustworthiness
consistent to the world knowledge

Seq2Set

Seq2Edits

Seq + Knowledge

Setups beyond standard Seq2seq are also important!
Comparing with Targets is Not Enough

A fire crew remains at Plasgran, Wimblington.

A fire has damaged a plastics factory in Cambridgeshire.

A fire crew remains at Plasgran, Oxfordshire.
Faithful to The Source

A fire crew remains at Plasgran, Wimblington.

A fire crew remains at Plasgran, Oxfordshire.

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

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

\[
P(z|x) = \prod_{t=1}^{z} P(z_t | y_1: j_{t-1}, \hat{z}_1: t-1, x_{k_t}, x)
\]

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

Datasets
(supervised, document - summary pairs)

WikiLarge & Small

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

newspaper

94,208/1129/1076

Models

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)
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

- **EditNTS** [Dong et al., ACL 19]
- **LaserTagger** [Malmi et al., EMNLP 19]
- **Levenshtein Transformer** [Gu et al., NeurIPS 19]
- **Fact preserving** [Iso et al., ACL 20]
- **EditBased Text Generation Models**
- **Our model**
- **Tensorflow version**
- **NAACL 22 tutorial**
Part II: Faithfulness

Takeaways:
Setups beyond **standard Seq2seq** are important for bounding the generation freedom (seq2edits)
Part III

Robustness

Learning beyond artifacts and biases

Faithfulness

consistent to the source

Trustworthiness

consistent to the world knowledge

Seq2Set

Seq2Edits

Seq + Knowledge

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.

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

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

External Knowledge!
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

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

E.g.
Location-based target entities in the dev. documents:
- 40% in the source
- 20% in the knowledge subgraph
- 40% neither
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 ...

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

**Entity Masking**
A large fire has broken out at a **[MASK]** in **[MASK]**...

**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**: Learning beyond artifacts and biases
- **Faithfulness**: Generation is consistent to the source
- **Trustworthiness**: Generation is consistent to the world 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

- **Robustness**
  - Seq2Set
    - Non-autoregressive models
  - RL, Loss & Optimization, Inductive priors, Multimodality, Adversarial attacks

- **Faithfulness**
  - Seq2Edits
    - Bounds by edit distances
  - Uncertainty, Scaling, Theoretical bounds

- **Trustworthiness**
  - Seq + Knowledge
    - Bounds by knowledge
  - 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
Language is for effective communication

Language is not suitable for complex thoughts

Language modeling
(Seq2seq: next word prediction)

Reasoning modeling
(Structure prediction beyond Seq2seq)

combine

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 Carenni
3. Text Generation with Text-Editing Models (NAACL 22)  
   Eric Malmi, Jonathan Mallinson, Aleksandr Chuklin, Jakub Adamek, Danil Mirylenka, Felix Stahlberg, Sebastian Krause, Shankar Kumar, Aliaksei Severyn