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)

c. Salient
   ■ contain important key points

d. Faithful
   ■ consistent with the source

e. Factual
   ■ consistent with the world knowledge

General requirements

Generated summary

Task-specific requirements

Salience
Faithfulness
Factuality

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?

---

**Paradigm**
- Generic pre-training
- Task adaptation

<table>
<thead>
<tr>
<th>Paradigm</th>
<th>Supervised learning</th>
<th>Transfer learning</th>
<th>Prompt-based learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic pre-training</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

---

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

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

- **Seq2Set** - control bias exploitation
- **Seq2Edit** - control hallucination
- **Seq + KB** - control with facts
Cooperation: Go big and Go Under Control

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.

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

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 $a$

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

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)$$

ROUGE: similarity between generated summary and gold-reference summary

$$ R(i, a) = \frac{1}{3} \sum_{k=1,2,L} \text{ROUGE-}k_f(i, a) $$
Structure of Policy \( p_\theta(\cdot|d) = \mu(\cdot|\pi_\theta(d)) \)

Deterministic \( \pi_\theta(d) \)

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

Salience Estimation

Encoder

document

\( \pi_i \in [0,1] \)

Decoder

\( T \)

\( T \)

Sampling w/o. replacement

Explore

Exploit

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

Dataset: CNN/DailyMail

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

- Comparable to seq2seq + RL [2]

---

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: 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

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**

![Diagram](image)

Complex Sent. $x$: Soldiers garrison the city

Minimum Edit Distance

$p(z|x)$

Expert Edit program $z$:

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

Simple sent. $y$: 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)
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

[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

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

Yue Dong, John Wieting and Pat Verga

Verify hallucination with world knowledge
Variants of hallucinations [1]

**Intrinsic:** generated text **contradicts source text**

**Extrinsic:** generated text is **not grounded in the source text**

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

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

<table>
<thead>
<tr>
<th>Location</th>
<th>Source Only</th>
<th>1 Hop</th>
<th>2 Hops</th>
<th>3 Hops</th>
</tr>
</thead>
<tbody>
<tr>
<td>XSUM</td>
<td>40.1%</td>
<td>59.8%</td>
<td>60.2%</td>
<td>60.3%</td>
</tr>
<tr>
<td>CNNDM$_{abs}$</td>
<td>52.3%</td>
<td>65.4%</td>
<td>66.1%</td>
<td>66.2%</td>
</tr>
</tbody>
</table>

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 ...

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...

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Abstractive</th>
<th>Extractive</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XSUM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>68.72</td>
<td>64.29</td>
<td>66.31</td>
</tr>
<tr>
<td>+ T5m</td>
<td>68.73</td>
<td>64.33</td>
<td>66.34</td>
</tr>
<tr>
<td>+ FILM</td>
<td><strong>73.40</strong></td>
<td><strong>65.32</strong></td>
<td><strong>71.60</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CNNNDM_{abs}</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5</td>
<td>29.58</td>
</tr>
<tr>
<td>+ T5m</td>
<td>28.95</td>
</tr>
<tr>
<td>+ FILM</td>
<td><strong>30.31</strong></td>
</tr>
</tbody>
</table>

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

---

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!

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

@YueDongCS
yue.dong@ucr.edu

Thanks to all my collaborators!

**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