Generalizable Factual Error Correction in Model Generated Summaries

Vidhisha Balachandran
18 May, 2023
NLP Today looks very different what it was 5 years ago!

Task-Specific Architectures

General, Multi-Task Architectures

Vidhisha Balachandran
They are pretrained on large, diverse sources of data

**DATA**
- Books
- Webpages
- Conversation Data
- Scientific Documents
- Code

**TASKS**
- Question Answering
- Sentiment Analysis
- Chatbots
- Summarization
- Information Extraction
They process unstructured text as sequence of tokens

23 Wall Street, also known as the [MASK] Building.

Robert Melancton Metcalfe is an American engineer and entrepreneur.

Human language has a rich, hierarchical structure.

Vidhisha Balachandran
They are pretrained on exponentially growing model sizes

https://textcortex.com/post/how-gpt-3-writing-tools-work
Pretrained LLMs can answer questions, reason and generate fluent text.

ChatGPT passes MBA exam given by a Wharton professor.

Scores of Stanford students used ChatGPT on final exams, survey suggests.

ChatGPT listed as author on research papers: many scientists disapprove.

Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach.

Meet Bard, Google’s Answer to ChatGPT.
Design Flaws - No transparency or control

Models not transparent by design
(Lipton, 2018; Vellido, 2020; Belinkov et al., 2020)

Models hard to control by design
(Ziegler et al., 2019; Dathathri et al., 2020)
Unintended effects due to such design flaws

Spurious correlations

Low generalizability

Factually Unreliable

The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.
My Research - Trustworthy NLP Systems

Designing large-scale, data-driven language generation models to be more **trustworthy, transparent and reliable** through **interpretable and controllable data and model designs**
Factual Unreliability of Model Generated Text

6/ It makes a TON of stuff up.

I decided to lightly fact check the response it gave to a journalist about picking a new phone.

The screenshot speaks for itself. I wouldn’t want this level of accuracy to make buying decisions.

Hello, this is Bing. I can help you find a phone with a good camera and battery life under $1500 AUD.

According to the web, some of the best camera phones in 2023 are:

- **Google Pixel 7 Pro**[^1]: It has a 50MP main camera, a 48MP telephoto camera with 4x optical zoom, and a 12MP ultra-wide camera. It also has a 12MP selfie camera and can record 8K video at 30fps. It has a 6.3-inch OLED display and a 5000mAh battery. It costs $999 USD[^2] which is about $704 AUD.

- **Samsung Galaxy S22 Ultra**[^3]: It has a 108MP main camera, a 12MP telephoto camera with 10x optical zoom, and a 12MP ultra-wide camera. It also has an 8MP selfie camera and can record 8K video at 24fps. It has a 6.8-inch AMOLED display and a 5000mAh battery. It costs $999 USD[^4] which is about $1408 AUD.

- **Apple iPhone 14 Pro**[^5]: It has a 12MP main camera, a 12MP telephoto camera with 3x optical zoom, and a 12MP ultra-wide camera. It also has a 12MP selfie camera and can record 8K video at 30fps. It has a 6.1-inch OLED display and a 3095mAh battery. It costs $999 USD[^6] which is about $1408 AUD.

These are some of the options you can consider. You can also check out the shopping results for more choices.

Google’s AI chatbot Bard makes factual error in first demo

/ The mistake highlights the biggest problem of using AI chatbots to replace search engines – they make stuff up.
The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking impact of the virus on communities.

As local data sources become less reliable, The Times will instead report information collected by the C.D.C. on its virus tracking pages.
The New York Times

**The Times Switches to C.D.C. Covid Data, Ending Daily Collection**

As local data sources become less reliable, The Times will instead report information collected by the C.D.C. on its virus tracking pages.

After more than three years of daily reporting on the number of Covid-19 cases and deaths in every county in the United States, The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking pages for the United States, only now they will be based on the latest information available from the federal government, not the Times's data set.

The tracking pages will still show data about hospital patients with Covid; reported cases and tests; and how many people have died from the virus. Data on vaccination rates and comparisons between vaccinated and unvaccinated populations will also remain.

A new interactive county map will show local levels of...
The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking ... ... impact of the virus on communities.

As local data sources become less reliable, The Times will stop reporting information collected by the C.D.C. on its pandemic headlines.
Interventions to Mitigate Factual Errors

Factuality Constraints and Objectives (Cao et al, 21)

Decoding Constraints (King et al, 22)

Summary w/ & w/o Factual Errors
Post-Editing to Correct Factual Errors
Challenges in Collecting Training Data

● Training Data: (Incorrect Summary, Correct Summary) Pairs

● Human Annotated Data
  ○ Expensive - Long Process to read and edit summaries
  ○ Subjective - Factuality is subjective

● Synthetic Data - Create adversarial data (incorrect summaries) using heuristic rules
  (Kryściński, et. al, 2020, Cao, et. al, 2020)
  ○ Low Generalizability - Only able to correct few heuristic errors trained on
  ○ Low Coverage - Hard to design heuristics for complex factual errors
FactEdit - Infilling LMs for Syntactic Adversarial Data Generation

Infilling based Generation of Adversarial Summaries

Seq-to-Seq Model
Factual Error Correction
[Vaccine for Covid-19]_{Subj} [is]_{Rel} [unlikely to be ready this year.]_{Obj}

[Vaccine for Ebola]_{Subj} [is]_{Rel} [unlikely to be ready this year.]_{Obj}

[Vaccine for Covid-19]_{Subj} [is]_{Rel} [under clinical trials.]_{Obj}

[Vaccine for Covid-19]_{Subj} [looks less]_{Rel} [unlikely to be ready this year.]_{Obj}
Infilling LMs for Candidate Generation

**Reference:** Vaccine for Covid-19 is unlikely to be ready this year.

**Context:** The first vaccine for Covid-19 ........ ready this year, although clinical trials have already started. For reference the vaccine for Ebola took ........

The first vaccine for [MASK] is ...... [SEP] clinical trials ...

Infilling Language Model

Covid-19
Infilling LMs for Candidate Generation

**Reference:** Vaccine for Covid-19 is unlikely to be ready this year.

Vaccine for [MASK] is unlikely to be ready this year, .... started. [SEP] For reference the vaccine for Ebola ...

**Context:** The first vaccine for Covid-19 ........ ready this year, although clinical trials have already started. For reference the vaccine for Ebola took ........

Coronavirus ✓

Covid-19 ✓

Ebola ✗

Polio ✗

Vaccine for Ebola is unlikely to be ready this year.

Vaccine for Polio is unlikely to be ready this year.

Beam-Search Candidates
Infilling LM - Training Data Generation

**Incorrect:** Vaccine for Ebola is unlikely to be ready this year.

**Correct:** Vaccine for Covid-19 is unlikely to be ready this year.

**Incorrect:** Vaccine for Covid-19 is being prepared for this year.

**Correct:** Vaccine for Covid-19 is unlikely to be ready this year.

**Incorrect:** Researchers expect to begin Phase 3 trials after production.

**Correct:** Researchers expect to begin production after Phase 3 trials.
Training Phrase-Level Error Correction

**Summary:** Vaccine for Ebola is unlikely to be ready this year.

**Context:** The first vaccine for Covid-19 ....... ready this year, although clinical trials have already started. For reference the vaccine for Ebola took .......

Vaccine for Ebola is unlikely to be ready this year. [SEP] The first vaccine for Covid-19 ....... ready this year, although clinical trials have already started. For reference the vaccine for Ebola took .......

Vaccine for Covid19 is unlikely to be ready this year.
Phrase-Level Error Correction

Incorrect Model Generated

Summary: World Leaders met Ban Ki Moon for UN Meeting in 2020.

Context: UN Sec. Gen. Antonio Gutteres met the leaders for … economic implications of global sanctions.

World Leaders met Ban Ki Moon for UN Meeting in 2020 [SEP]
Pandemic Response … [SEP] UN Sec. Gen. Antonio Gutteres met the leaders for …

Seq2Seq Model

World Leaders met Antonio Gutteres for UN Meeting in 2020.
Data & Experiment Settings

- **Standard Datasets - CNN/DM + XSum**
  - Full Test Set evaluation using BART Model

- **FRANK Benchmark - human annotated judgements for error categorization**
  - Generated outputs from 5 models for CNN/DM
  - Generated outputs from 4 models for XSum

- **Evaluation**
  - Rouge - fluency, grammar and content
  - FactCC - semantic, sentence level errors
  - Ent-DAE - discourse errors
Post-Editing factual errors improves factuality of generated summaries

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
<th>FactCC</th>
<th>Ent-DAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN/DM</td>
<td>Bart (Lewis et al., 2020)</td>
<td>44.07</td>
<td>21.08</td>
<td>41.01</td>
<td>75.78</td>
<td>74.85</td>
</tr>
<tr>
<td></td>
<td>Cao (Cao et al., 2020)</td>
<td>42.72</td>
<td>20.59</td>
<td>39.92</td>
<td>49.98</td>
<td>74.83</td>
</tr>
<tr>
<td></td>
<td>FACTEDIT</td>
<td>42.17</td>
<td>20.22</td>
<td>39.37</td>
<td>75.49</td>
<td><strong>75.71</strong></td>
</tr>
<tr>
<td></td>
<td>FACTEDIT + FactCC Filter (FF)</td>
<td>42.53</td>
<td>20.48</td>
<td>39.74</td>
<td><strong>76.03</strong></td>
<td>75.36</td>
</tr>
</tbody>
</table>
Post-Editing factual errors improves factuality of generated summaries

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
<th>FactCC</th>
<th>Ent-DAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN/DM</td>
<td>Bart (Lewis et al., 2020)</td>
<td>44.07</td>
<td>21.08</td>
<td>41.01</td>
<td>75.78</td>
<td>74.85</td>
</tr>
<tr>
<td></td>
<td>Cao (Cao et al., 2020)</td>
<td>42.72</td>
<td>20.59</td>
<td>39.92</td>
<td>49.98</td>
<td>74.83</td>
</tr>
<tr>
<td></td>
<td>FACTEDIT</td>
<td>42.17</td>
<td>20.22</td>
<td>39.37</td>
<td>75.49</td>
<td>75.71</td>
</tr>
<tr>
<td></td>
<td>FACTEDIT + FactCC Filter (FF)</td>
<td>42.53</td>
<td>20.48</td>
<td>39.74</td>
<td><strong>76.03</strong></td>
<td>75.36</td>
</tr>
<tr>
<td>XSum</td>
<td>Bart (Lewis et al., 2020)</td>
<td>34.71</td>
<td>15.04</td>
<td>27.40</td>
<td>21.93</td>
<td>20.03</td>
</tr>
<tr>
<td></td>
<td>Cao (Cao et al., 2020)</td>
<td>33.64</td>
<td>14.71</td>
<td>26.49</td>
<td>7.01</td>
<td>20.03</td>
</tr>
<tr>
<td></td>
<td>FACTEDIT</td>
<td>33.58</td>
<td>14.68</td>
<td>26.71</td>
<td><strong>23.91</strong></td>
<td>20.13</td>
</tr>
<tr>
<td></td>
<td>FACTEDIT + FactCC Filter (FF)</td>
<td>33.58</td>
<td>14.68</td>
<td>26.71</td>
<td><strong>23.91</strong></td>
<td>20.13</td>
</tr>
</tbody>
</table>
Infilling based generated data generalizes correction across models

FactEdit improves performance across most data and model settings

Large Gains in abstractive XSum Dataset
Infilling based generated data generalizes correction across error types

FactEdit improves performance across error types Especially good at correcting semantic frames.
Infilling based generated data generalizes correction across datasets

Performance across Models on XSum

FactEdit trained on data from longer summaries (CNN/DM) improves results on headline generation (XSum)

Vidhisha Balachandran
Summary

● Data Generation leveraging **Infilling LMs to edit fine-grained phrases**
  ○ Editing at syntactic units provides control over data generation process
  ○ Infilling LMs provide challenging, generalizable generation process

● **FactEdit** - **Phrase-level Factual Error Correction**
  ○ Syntactic Phrase based training data enables fine-grained correction
  ○ Correct Factual Errors without modifying other generated text

● **Generalizable** across models and error categories
  ○ Generated training data includes diverse examples of errors
  ○ Flexible and Diverse data generation generalizes error correction across multiple models and error categories
Thank you!

- Data Generation leveraging **Infilling LMs to edit fine-grained syntactic units**
  - Editing at syntactic units provides control over data generation process
  - Infilling LMs provide challenging, generalizable generation process

- **FactEdit - Phrase-level Factual Error Correction**
  - Syntactic Phrase based training data enables fine-grained correction
  - Correct Factual Errors without modifying other generated text

- **Generalizable** across models and error categories
  - Generated training data includes diverse examples of errors
  - Flexible and Diverse data generation generalizes error correction across multiple models and error categories

Questions: vbalacha@cs.cmu.edu

![QR Code]

Paper  Code