Human-AI Collaboration for Neural Text Generation with Interpretable Neural Networks

Sebastian Gehrmann

Thesis Defense

Committee Members
Barbara Grosz
Sasha Rush
Stuart Shieber

Oct 18, 2019
This is Jesse, a journalist. Jesse has a ton of work.
Maybe AI can help reduce the workload?
Introduce AI-reen, a text-generation model.
Jesse could give some of the workload to AI-reen. Doing so, Jesse would give up her agency over that work.
But Al-reen is biased and makes mistakes! Jesse still needs to provide oversight over its work.
By **collaborating** with Al-reen, Jesse could gain the benefits of automation without losing her agency.
Problem

Explain Suggestion

Provide Feedback

Update Suggestion

Accept

Accepted solution
They want to collaboratively summarize a document.
Both have an idea how to summarize it.
If AI-reen was human, it could communicate its reasoning. But its prediction are not **interpretable**.
Even if it could explain its suggestion, it can’t incorporate feedback from Jesse.

I picked this phrase, because…

I don’t like it.
Interpretability is necessary, but we also need **controllability**.

I picked this phrase, because…

How about … instead?
Let’s empower humans to collaborate with AI!

++

Summarization [EMNLP ’18]
Data2Text [INLG ’18]
Section Title Generation [NAACL ’19]
TL;DR Generation [INLG ’19]

LSTMVis [InfoVis ’17]
Phenotyping Saliency [PloS one, ’17]
Seq2Seq-Vis [VAST ’18]
Model Selection [DeepStruct ’19]
Modeling Capacity [Formal Languages ’19]

Collaborative Semantic Inference [VAST ’19]
Detecting Fake Text with GLTR [ACL Demo ’19]
Automated Mediation [Behavior & Technology ’19]
Outline

1. Background: Sequence Modeling for NLP
2. Incorporating Content Selection into a Summarization Model
3. How to Understand Predictions?
4. Collaborating with the Model to Summarize
The small dog owns a yellow ball.

$p(y_{t+1} \mid y_1, \ldots, y_t)$
The small dog owns a yellow ball.

[Elman '90, Hochreiter & Schmidhuber '97]
The small dog owns a yellow ball.

[Elman ‘90, Hochreiter & Schmidhuber ‘97]
The small dog owns a yellow ball.

\[ p \]

large small child dog

\[ y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad y_6 \quad y_7 \quad y_8 \]
The small dog owns a yellow ball.

\[ p(y_3 \mid y_1, y_2, x) \]

\[ p(\text{next word} \mid \text{Der kleine, The small dog...}) \]
The small dog owns a yellow ball.

Encoder

Decoder

[Bahdanau et al. '14, Sutskever et al. '14]
Attention $p(a_t | x, y_{1:t})$

The small dog owns a yellow ball.

Encoder

Decoder

[Bahdanau et al. ‘14, Sutskever et al. ‘14]
The small dog owns a yellow ball.

Context: \[ \sum_{s=1}^{S} a_t^s x_s \]

Encoder

Decoder

[ Bahdanau et al. ’14, Sutskever et al. ’14 ]
The small dog owns a yellow ball.

[Bahdanau et al. ‘14, Sutskever et al. ‘14]
Consider an **abstractive summarization** problem, with

**Input** \( x_1, \ldots, x_S \)

**Summary** \( y_1, \ldots, y_T \)

Train a summarizer to maximize \( p(y \mid x) \).

[Gehrmann, Deng, and Rush, EMNLP ’18]
Attention $p(a_t | x, y_{1:t})$

Encoder

Decoder

[Vinyals et al. ‘15, Filippova et al. ‘15, Gu et al. ’16, See et al. ‘17]
The **copy mechanism** uses a binary soft switch $z_t$ that determines whether the model copies or generates.

\[
p(y_{t+1} | x, y_{1:t}) = p(\checkmark \bigcirc | x, y_{1:t}) \times \
+ p(\bigcirc \times | x, y_{1:t}) \times p
\]
The **copy mechanism** uses a binary soft switch $z_t$ that determines whether the model copies or generates.

\[
p(y_{t+1} \mid x, y_{1:t}) = p(z_t=1 \mid x, y_{1:t}) \times p(y_{t+1} \mid z_t=1, x, y_{1:t}) + p(z_t=0 \mid x, y_{1:t}) \times p(y_{t+1} \mid z_t=0, x, y_{1:t})
\]

\[
\sigma(W h_t + b)
\]

Reusing $p(a_t \mid x, y_{1:t})$

\[
1 - \sigma(W h_t + b)
\]

Standard model prediction
Just because a model can copy, should it?
Abstractive summarizers over-extract.

“Angela Merkel and her husband, chemistry professor Joachim Sauer, are spotted on their annual Easter trip to the island of Ischia, near Naples.”
The model fails at content selection!

Consider the content selection as **word-level extractive summarization**.

Let \( t_1, \ldots, t_S \) denote a binary indicator whether a source word is used in a summary.

Train a model to maximize \( p(t \mid x) \).
How to generate supervised data?

The small dog owns a large yellow ball.
The big dog from next door chases the ball.

Big dog chases small dog's ball.
The small dog owns a large yellow ball.
The big dog from next door chases the ball.

Big dog chases small dog's ball.
The small dog owns a large yellow ball.
The big dog from next door chases the ball.

Big dog chases small dog’s ball.
The small dog owns a large yellow ball.
The big dog from next door chases the ball.

Big dog chases small dog's ball.
The small dog owns a large yellow ball.
The big dog from next door chases the ball.

Big dog chases small dog's ball.
The small dog owns a large yellow ball.
The big dog from next door chases the ball.

Big dog chases small dog’s ball.
The small dog owns a large yellow ball. The big dog from next door chases the ball.

Big dog chases small dog's ball.
The small dog owns a large yellow ball.
The big dog from next door chases the ball.

Content Selector Model based on ELMo
Control copied content with Bottom-Up Attention by restricting what can be copied to important content.
Control copied content with Bottom-Up Attention by restricting what can be copied to important content.

Let $q_s$ denote the selection probability from the content selector. Let $\epsilon$ denote an importance threshold.

Modify the copy-attention such that

$$p(\tilde{a}_t^s \mid x, y_{1:t}) = \begin{cases} p(a_t^s \mid x, y_{1:t}) & q_s > \epsilon \\ 0 & \text{ow.} \end{cases}$$
Bottom-Up Attention $p(\tilde{a}_t | x, y_{1:t})$
The improvements were consistent across two evaluated datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointer-Generator (See et al., 2017)</td>
<td>36.4</td>
<td>15.7</td>
<td>33.4</td>
</tr>
<tr>
<td>Pointer-Generator + Coverage (See et al., 2017)</td>
<td>39.5</td>
<td>17.3</td>
<td>36.4</td>
</tr>
<tr>
<td>Pointer-Generator (our implementation)</td>
<td>36.3</td>
<td>16.2</td>
<td>33.4</td>
</tr>
<tr>
<td>Pointer-Generator + Coverage Penalty</td>
<td>39.1</td>
<td>17.4</td>
<td>36.1</td>
</tr>
<tr>
<td>CopyTransformer + Coverage Penalty</td>
<td>39.3</td>
<td>17.5</td>
<td>36.5</td>
</tr>
<tr>
<td>Bottom-Up Summarization</td>
<td>41.2</td>
<td>18.7</td>
<td>38.3</td>
</tr>
<tr>
<td>Bottom-Up Summarization (CopyTransformer)</td>
<td>41.0</td>
<td>18.4</td>
<td>38.2</td>
</tr>
</tbody>
</table>
“Angela Merkel and her husband, chemistry professor Joachim Sauer, are spotted on their annual Easter trip to the island of Ischia, near Naples.”

“Angela Merkel and her husband are spotted on their Easter trip.”
There is still work to be done…

Copy actions of different lengths [%]

Reference
Pointer-Generator
Bottom-Up Attention

1 2 3-5 6-10 11+
Summarization models struggle in real-world scenarios!

How do we make the generation of a summary **collaborative**?
The Users of Interpretability and Collaboration

Architect
Trainer
End User

[Strobelt*, Gehrmann, et al., InfoVis ‘17]
The Target of Interpretability and Collaboration

\[ \theta \]

Model

\[ \hat{y} \]

Decision

[Gehrmann*, Strobelt*, et al., VAST ’19]
The Coupling of Model and Interface

- Passive Observation
- Interactive Observation
- Interactive Collaboration

[Gehrmann*, Strobelt*, et al., VAST ’19]
die längsten reisen fangen an, wenn es auf den straßen dunkel wird.

(a) Passive Observation
(b) Interactive Observation
(c) Interactive Collaboration

[Strobelt*, Gehrmann*, et al., VAST ’18]
Collaborative Summarization?
Collaborative Summarization!
\[ p(y \mid x) \quad \text{Train-Endpoint predictor} \]

\[ p(z \mid x) \]

\[ p(z \mid x, y) \]
Train-Endpoint + Path predictor

\[ p(y, z \mid x) \]

\[ p(z \mid x) \]

\[ p(z \mid x, y) \]
前进

$p(z_{\text{left}} = 0)$
Scientists at NASA are one step closer to understanding how much water could have existed on primeval Mars.

These new findings also indicate how primitive water reservoirs there could have evolved over billions of years, indicating that early oceans on the Red Planet might have held more water than Earth's Arctic Ocean, NASA scientists reveal in a study published Friday in the journal Science.
Scientists at NASA are one step closer to understanding how much water could have existed on the primeval Mars.

These new findings also indicate how primitive water reservoirs there could have evolved over billions of years, indicating that early oceans on the red planet might have held more water than Earth's arctic ocean, NASA scientists reveal in a study published Friday in the journal Science.

"Our study provides a solid estimate of how much water Mars once had, by determining how much water was lost to space," said Gerónimo Villanueva, a scientist at NASA's Goddard Space Flight Center.

"With this work, we can better understand the history of water on Mars.

"To find answers to this age-old question about Martian water molecules, scientists used the world's three major infrared telescopes, in Chile and Hawaii, to measure traces of water in the planet's atmosphere over a range of areas and seasons, spanning from March 2008 to January 2014,"

"From the ground, we could take a snapshot of the whole hemisphere on a single night," said Goddard's Michael Mumma.

Scientists looked at the ratio of two different forms – or isotopes – of water, \( H_2O \) and \( HDO \),

the latter is made heavier by one of its hydrogen atoms, called deuterium, which has a neutron at its core in addition to the proton that all hydrogen atoms have. That weighed down HDO more, while larger amounts of hydrogen from \( H_2O \) floated into the atmosphere, broke away from Mars' low gravity and disappeared into space.
Blue: What do I want to use?
Red: What has been summarized

Where did it come from?
```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
from mars -- perhaps when an asteroid hit -- and landed on earth some 4.5 billion years ago.

they were able to determine how much that ratio had changed over time and estimate how much water has disappeared from mars -- about 87 %.

the findings indicate that the red planet could have had its fair share of blue waters, possibly even yielding an ocean.

according to nasa, there might have been enough water to cover up to 20 % of mars' surface.

that would amount to an ocean proportionally larger than the atlantic on earth.

"this ocean had a maximum depth of around 5,000 feet or around one mile deep," said villanueva.

nasa scientists say that much of this water loss happened over billions of years, along with a loss of atmosphere.

and as the planet's atmospheric pressure dropped, it was harder for water to stay in liquid form.

heat also contributed to its evaporation.

as a result, the remaining primeval ocean water continued to move toward the poles, where it eventually froze.

"with mars losing that much water, the planet was very likely wet for a longer period of time than was previously thought, suggesting it might have been habitable for longer," said mumma.

cnn's ben brunfield contributed to this report.
<table>
<thead>
<tr>
<th>Sentence</th>
<th>Edit</th>
<th>Delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;t&gt; Primitive water reservoirs there could have evolved over billions of years, scientists say. &lt;/t&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;t&gt; Early oceans on the red planet might have held more water than Earth's Arctic Ocean. &lt;/t&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;t&gt; Scientists used the world's three major infrared telescopes to measure traces of water in the planet's atmosphere over a range of areas and seasons. &lt;/t&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;t&gt; Water trapped in Mars' polar ice caps has a much higher level of HDO than fluid water on earth. &lt;/t&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the water is ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
primitive water reservoirs there could have evolved over billions of years, scientists say.

early oceans on the red planet might have held more water than earth's arctic ocean.

scientists used the world's three major infrared telescopes to measure traces of water in the planet's atmosphere over a range of areas and seasons.

water trapped in mars' polar ice caps has a much higher level of hdo than fluid water on earth.

the water is very likely wet for a longer period of time than was previously thought, the scientists say.
From Mars -- perhaps when an asteroid hit -- and landed on Earth some 4.5 billion years ago, they were able to determine how much that ratio had changed over time and estimate how much water has disappeared from Mars -- about 87%.

The findings indicate that the red planet could have had its fair share of blue waters, possibly even yielding an ocean.

According to NASA, there might have been enough water to cover up to 20% of Mars' surface.

That would amount to an ocean proportionally larger than the Atlantic Ocean.

This ocean had a maximum depth of around 5,000 feet or around one mile deep, said Villanueva.

NASA scientists say that much of this water loss happened over billions of years, along with a loss of atmosphere.

As the planet's atmospheric pressure dropped, it was harder for water to stay in liquid form.

Heat also contributed to its evaporation.

As a result, the remaining primeval ocean water continued to move toward the poles, where it eventually froze.

"With Mars losing that much water, the planet was very likely wet for a longer period of time than was previously thought, suggesting it might have been habitable for longer," said Mumma.

Cnn's Ben Brumfield contributed to this report.
CSI can make models collaborative

Leveraging the underlying structure of a problem can lead to better neural models. We can add interpretable and controllable latent variables that follow this structure. Exposing the variables in an interface allows end users to manipulate the model reasoning process.

This allows end users to retain agency over the automated process.
Summarization

Content Selection  Bottom-Up Attention

Structured Summarization

Source  Masked Source  Summary

Data2Text

TL;DR Generation

Section Title Generation
(a) Sampling from language models yields a detectable signature. See, for example, how this owl was not predicted. However, generated language is mostly green and yellow, see here: (generated) Raspberry Pi is a tiny computer which comes with an ARM processor, so it’s easy to hack. But now the makers of this tiny computer have come up with an even more powerful chip: the Raspberry Pi 2.
What are future opportunities?
NLP is exciting at the moment... and scary.
How do we use large LM’s for controlled generation?

Captioning  Data2Text  Abstractive Summarization

GPT-2

Ctrl
How can we detect and prevent biases in learned representation that influence downstream tasks?
How do we evaluate generated natural language at a time where the content, and not lexical overlap, matters?

How can we measure its affect on humans?
Acknowledgements <3
Acknowledgements <3

Many thanks to…
• My Twitter followers
• All members of the Harvard HCI and NLP groups
• My friends and roommates
• My American family
• My German family
We can achieve Human-AI collaboration through interpretable and controllable neural models.