Dissecting Generation Modes for Abstractive Summarization Models via Ablation and Attribution

Jiacheng Xu and Greg Durrett
The University of Texas at Austin
Well-established techniques for interpreting classification/NLU models

(Ribeiro et al., 2016; Wiegrefe & Pinter, 2019)
Interpreting Summarization Models

Well-established techniques for interpreting classification/NLU models

(Ribeiro et al., 2016; Wiegreffe & Pinter, 2019)

How to interpret complicated sequential decisions?
Overview

Our contribution: a two-stage decision interpretation framework for summarization

For each time step: (1) Does the model need input context? (2) If yes, which input tokens matter?
Overview

Our contribution: a two-stage decision interpretation framework for summarization

For each time step: (1) Does the model need input context? (2) If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?
• in 20% of cases our model functions as a LM
Overview

Our contribution: a two-stage decision interpretation framework for summarization

For each time step: (1) Does the model need input context? (2) If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?
- in 20% of cases our model functions as a LM

Attribution: what part of the input leads to the decision?
- many attribution methods; what works best?
- we propose a framework for evaluation
Overview

Our contribution: a two-stage decision interpretation framework for summarization

For each time step:
1. Does the model need input context?
2. If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?
• in 20% of cases our model functions as a LM

Attribution: what part of the input leads to the decision?
• many attribution methods; what works best?
• we propose a framework for evaluation
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival […]

For each time step, we provide input and prefix, and the model predicts the next token.

David Cameron has […]
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival […]

For each time step, we provide input and prefix, and the model predicts the next token.
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival […]

Prefix
David Cameron has urged Londoners to vote
Input Article
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival […]

Prefix
David Cameron has urged Londoners to vote

Model
p( for | , , prefix) = 0.96

Model w/o Input Article
p( for | , , prefix) = 0.95

Human: Why does the model say “for”?  
Model: I am confident! Do my ablated versions agree with me?  
Agree!
# Ablation of

<table>
<thead>
<tr>
<th>Input</th>
<th>Model</th>
<th>Question</th>
<th>Ablation Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>🎓📝</td>
<td>Sum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>🎓❌📝</td>
<td>Sum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>🎓❌📝</td>
<td>LM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ablation of Do we need context?

<table>
<thead>
<tr>
<th>Input</th>
<th>Model</th>
<th>Question</th>
<th>Ablation Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>🧡💡📝</td>
<td>Sum</td>
<td>Do we need context?</td>
<td></td>
</tr>
<tr>
<td>✿🚫📝</td>
<td>Sum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✿🚫📝</td>
<td>LM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Ablation of

<table>
<thead>
<tr>
<th>Input</th>
<th>Model</th>
<th>Question</th>
<th>Ablation Target</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sum</td>
<td>Do we need context?</td>
<td>How much the decoder-only summarization model (Sum) agrees with full model (LM)?</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ablation of

<table>
<thead>
<tr>
<th>Input</th>
<th>Model</th>
<th>Question</th>
<th>Ablation Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>Sum</td>
<td>Do we need context?</td>
<td>= How much the decoder-only summarization model ((\hat{\text{Sum}})) agrees with full model ((\hat{\text{LM}}))?</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>Do we even need summarization fine-tuning?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ablation of

<table>
<thead>
<tr>
<th>Input</th>
<th>Model</th>
<th>Question</th>
<th>Ablation Target</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sum</td>
<td>Do we need context?</td>
<td>How much the decoder-only summarization model (coffee) agrees with full model (hat)</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td>Do we even need summarization fine-tuning?</td>
<td>How much the generic LM (coffee) agrees with full model (hat)</td>
</tr>
</tbody>
</table>
Do we even need summarization fine-tuning?

How much the decoder-only summarization model (Sum) agrees with full model (LM)?

How much the generic LM (LM) agrees with full model (LM)?

Ablation of `Do we need context?`

How much the generic LM (LM) agrees with full model (LM)?

Axis on Map:

- $x$
- $y$
Distance function

\[ d(p, q) = \sum_{i} |p_i - q_i| \]
Distance function

\[
d(p, q) = \sum_{i} |p_i - q_i|
\]

Context

Frequency: ~70%
Ablated models are distant from full model; Input is needed.

built in 1993 by James
Mapping Generation Modes

Frequency: ~20%
All ablated models agree with full model; Input is not needed.

According to Barack Obama
Mapping Generation Modes

Frequency: ~20%
All ablated models agree with full model; Input is not needed.

According to Barack Obama
Memorization Bias

Map on XSum
with Generation Modes

FT
LM
PT

Context

\( d(\alpha, \beta) \)

\( p \)
Memorization Bias

Pre-Training bias
Frequency: ~2%
Fine-tuning a LM on training data causes it to work less well.

Example
Gail Scott was desperate to emulate Kylie __?__

- $p(\text{Jenner}) = 0.90$
- $p(\text{Minogue}) = 0.80$
- $p(\text{Jenner}) = 0.99$

Map on XSum with Generation Modes

Pre-Training bias
Frequency: ~2%
Fine-tuning a LM on training data causes it to work less well.
Memorization Bias

Pre-Training bias
Frequency: ~2%
Fine-tuning a LM on training data causes it to work less well.

Example
Gail Scott was desperate to emulate Kylie __?__

\[
\begin{align*}
    p(\text{Jenner}) &= 0.90 \\
    p(\text{Minogue}) &= 0.80 \\
    p(\text{Jenner}) &= 0.99
\end{align*}
\]

Fine-Training bias
Frequency: ~2%
Fine-tuned decoder-only model without input is a close match but the pre-trained LM is not.

Example
In our series of letters from African journalists, […]

0.5% of ref summaries in XSum
Memorization Bias

**Pre-Training bias**
- Frequency: ~2%
- Fine-tuning a LM on training data causes it to work less well.

  Example
  Gail Scott was desperate to emulate Kylie __p__

  \[ p(\text{Jenner}) = 0.90 \]

  \[ p(\text{Minogue}) = 0.80 \]

  \[ p(\text{Jenner}) = 0.99 \]

**Fine-Tuning bias**
- Frequency: ~2%
- Fine-tuned decoder-only model without input is a close match but the pre-trained LM is not.

  Example
  In our series of letters from African journalists, […]

  0.5% of ref summaries in XSum
Overview

Our contribution: a two-stage decision interpretation framework for summarization

For each time step: (1) Does the model need input context? (2) If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?
- in 20% of cases our model functions as a LM

Attribution: what part of the input leads to the decision?
- many attribution methods; what works best?
- we propose a framework for evaluation
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]

Prefix
David

BART

Cameron
Find the context which actually matters!

Methods: gradient, occlusion, etc.

Human: Why does the model say “Cameron”?

Model: Ablated version disagrees → Input matters.

Human: So what exactly do you look at?

Input Article
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival […]
Attribution

Challenge

• hard to compare highlights
• inconsistency among different attribution methods

What we want

an evaluation protocol for attributions

Integrated Gradient

Attention
Evaluation Protocol
Assign score for each token according to attribution
Evaluation Protocol

Assign score for each token according to attribution

Sort them according to the score and select top-k

<table>
<thead>
<tr>
<th>Rank</th>
<th>Token</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>prime</td>
<td>0.36</td>
</tr>
<tr>
<td>2</td>
<td>minister</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>warned</td>
<td>0.17</td>
</tr>
</tbody>
</table>

... the prime minister warned David Cameron
Evaluation Protocol

Assign score for each token according to attribution

Sort them according to the score and select top-k

Does the selected set of tokens help recover the prediction?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Token</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>prime</td>
<td>0.36</td>
</tr>
<tr>
<td>2</td>
<td>minister</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>warned</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Input: ...the **prime minister** warned...  
Prefix: prime minister +  
$p(Cameron)$: 0.97

Input: he talked to  
Prefix:  
$p(Cameron)$: 0.08
Conclusion

Why do we do ablation before attribution?
• It identifies generation modes and allow us to deploy different tools on each mode.

Can you extend the framework to other NLG tasks?
• Yes!

How effective and accurate are attribution methods?
• Fine for many cases, but still a long way to go.