Text Generation with Exemplar-based Adaptive Decoding

Hao Peng, Ankur Parikh, Manaal Faruqui, Bhuwan Dhingra, Dipanjan Das

@NAACL
June 4, 2019
Outline

❖ Background and Overview
❖ Adaptive Decoding
❖ Experiments
A Portuguese train derailed in Oporto on Wednesday, killing three people.

Portuguese train derailed, killing three.
A Portuguese train derailed in Oporto on Wednesday, killing three people.

Portuguese train derailed, killing three.
Exemplar-informed Generation

Source X \[\xrightarrow{\text{Retrieve}}\] Exemplar Z

A Portuguese train derailed in Oporto on Wednesday, killing three people

Two die in a Britain train collision

Goal y Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
Exemplar-informed Generation

What to say
Source $X$  Retrieve Training target  How to say it
Exemplar $Z$

A Portuguese train derailed in Oporto on Wednesday, killing three people

Two die in a Britain train collision

Motivation

• Better performance
  Cao et al., 2018; Zhang et al., 2018

• Diversity and interpretability
  Guu et al., 2017; Wiseman et al., 2018

Goal $y$  Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
Exemplar-informed Generation

**What to say**

Source \( X \)

Retrieved training target

**How to say it**

Exemplar \( Z \)

A Portuguese train derailed in Oporto on Wednesday, killing three people

Two die in a Britain train collision

**Training pairs**

\((x_1, y_1)\)

\((x_2, y_2)\)

\((x_3, y_3)\)

\((x_4, y_4)\)

\(\vdots\)

**Goal** \( y \)

Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
Exemplar-informed Generation

What to say
Source X

Retrieve
Training target

How to say it
Exemplar Z

Training pairs

(\(x_1, y_1\))
(\(x_2, y_2\))
(\(x_3, y_3\))
(\(x_4, y_4\))

A Portuguese train derailed in Oporto on Wednesday, killing three people

Two die in a Britain train collision

Goal y
Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
A Portuguese train derailed in Oporto on Wednesday, killing three people.

Two die in a Britain train collision.

What to say
Source X

How to say it
Exemplar Z

Retrieve
Training target

Similar

Training pairs

Goal y
Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
Status Quo

A Portuguese train derailed in Oporto on Wednesday, killing three people. Two die in a Britain train collision.

\[ \text{Enc} ([x; z]) \]

Goal \( y \)  Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
A Portuguese train derailed in Oporto on Wednesday, killing three people. Two die in a Britain train collision.

\[
\text{Dec} \left( y \mid \text{Enc} \left( [x; z] \right) \right)
\]

Goal \( y \) Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
Status Quo

A Portuguese train derailed in Oporto on Wednesday, killing three people. Two die in a Britain train collision.

**What to say**

\[ \text{Enc} \left( [x; z] \right) \]

**How to say it**

\[ \text{Dec} \left( y \mid \text{Enc} \left( [x; z] \right) \right) \]

**Goal**

Three die in a Portuguese train derailment

Guu et al., 2017; Cao et al., 2018
A Portuguese train derailed in Oporto on Wednesday, killing three people. Two die in a Britain train collision.

Guu et al., 2017; Cao et al., 2018
Overview

Motivation
• Encoder $\longleftrightarrow$ what to say
• Decoder $\longleftrightarrow$ how to say it

Method: Adaptive Decoding
• Exemplar-specific decoder

\[
P(y \mid \text{Source } X, \text{Exemplar } Z)
\]
Overview

Motivation
• Encoder $\leftrightarrow$ *what to say*
• Decoder $\leftrightarrow$ *how to say it*

Method: Adaptive Decoding
• Exemplar-specific decoder
• Drop-in replacement in seq2seq
Overview

Motivation
• Encoder $\leftrightarrow$ what to say
• Decoder $\leftrightarrow$ how to say it

Method: Adaptive Decoding
• Exemplar-specific decoder
• Drop-in replacement in seq2seq

Experiments
• Summarization
• Data2text generation
Outline

- Background and Overview
- Adaptive Decoding
- Experiments
Adaptive Decoder

Goal

• Customized decoder for each exemplar.
Adaptive Decoder

Goal
• Customized decoder for each exemplar.

Key Points
• Exemplar-informed interpolation of backbones.
Adaptive Decoder

Goal
• Customized decoder for each exemplar.

Key Points
• Exemplar-informed interpolation of backbones.
• Low-rank constraints by construction.
Adaptive Decoder

\[(\mathbf{W}_1, \mathbf{c}_1), (\mathbf{W}_2, \mathbf{c}_2), (\mathbf{W}_3, \mathbf{c}_3)\]
Adaptive Decoder

\[
\begin{align*}
&W_1, c_1 \\
&W_2, c_2 \\
&W_3, c_3
\end{align*}
\]

\[
\text{AdaDec}_z = \sigma_1 W_1 + \sigma_2 W_2 + \sigma_3 W_3
\]
Adaptive Decoder

\[
W = \sigma_1 W_1 + \sigma_2 W_2 + \sigma_3 W_3
\]
Adaptive Decoder

\[ W_1, c_1 \]

\[ W_2, c_2 \]

\[ W_3, c_3 \]

\[ p = \text{RNN}(z) \]
Adaptive Decoder

\[
p = RNN(z)
\]

\[
\begin{bmatrix}
\sigma_1 \\
\sigma_2 \\
\sigma_3
\end{bmatrix} = \begin{bmatrix}
p^\top c_1 \\
p^\top c_2 \\
p^\top c_3
\end{bmatrix}
\]

\[
W = \sigma_1 W_1 + \sigma_2 W_2 + \sigma_3 W_3
\]
Low-rank Constraints

\[ W = \sigma_1 W_1 + \sigma_2 W_2 + \sigma_3 W_3 \]
Low-rank Constraints

\[ W = \sigma_1 W_1 + \sigma_2 W_2 + \sigma_3 W_3 \]

Too many params!
Low-rank Constraints

\[
W = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\top + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^\top
\]
Low-rank Constraints

\[ W = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \sigma_3 u_3 v_3^\top \]

\( W \) = \( \sigma_1 \) \( u_1 v_1^\top \) + \( \sigma_2 \) \( u_2 v_2^\top \) + \( \sigma_3 \) \( u_3 v_3^\top \)

\( W = \sigma_1 \) \( u_1 v_1^\top \) + \( \sigma_2 \) \( u_2 v_2^\top \) + \( \sigma_3 \) \( u_3 v_3^\top \)

Rank \( \leq 3 \)

Rank = 1
Low-rank Constraints

\[ W = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \sigma_3 u_3 v_3^\top + \cdots \]

\[ W = \sigma_1 + \sigma_2 + \sigma_3 + \cdots \]

\[ \text{Rank} \leq d \]

\[ \text{Rank} = 1 \]

\[ d \]
Low-rank Constraints

\[ W = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top + \sigma_3 u_3 v_3^\top + \cdots \]

\[ W = \sigma_1 \underbrace{\cdots} + \sigma_2 \underbrace{\cdots} + \sigma_3 \underbrace{\cdots} + \cdots \]

\( \text{Rank} \leq d \)

\( \text{Rank} = 1 \)

\( d \)

\( \mathcal{O}(d^3) \rightarrow \mathcal{O}(d^2) \)
Walkthrough

Source $X$ \xrightarrow{\text{Retrieve training target}} \text{Exemplar } Z
Walkthrough

Source $X$ \xrightarrow{\text{Retrieve Training target}} \text{Exemplar } z

$p = \text{RNN}(z)$

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix} = \begin{bmatrix} p^\top c_1 \\ p^\top c_2 \\ p^\top c_3 \end{bmatrix}$$
Walkthrough

Source $X$ \(\xrightarrow{\text{Retrieve}}\) Training target \(\xrightarrow{}\) Exemplar $Z$

\[ p = \text{RNN}(z) \]

\[
\begin{bmatrix}
\sigma_1 \\
\sigma_2 \\
\sigma_3
\end{bmatrix} = 
\begin{bmatrix}
p^\top c_1 \\
p^\top c_2 \\
p^\top c_3
\end{bmatrix}
\]
Walkthrough

Source $X$ \xrightarrow{\text{Retrieve Training target}} \text{Exemplar } Z

$$p = \text{RNN}(z)$$

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix} = \begin{bmatrix} p^\top c_1 \\ p^\top c_2 \\ p^\top c_3 \end{bmatrix}$$

Diagram:

- **Enc**
- **AdaDec**

\[ \cdots \]
Outline

❖ Background and Overview
❖ Adaptive Decoding
❖ Experiments
Experiments: Summarization

Datasets:

• Gigaword. Rush et al., 2015
• New York Times (NYT). Durrett et al., 2016
Experiments: Summarization

Datasets:
- Gigaword. Rush et al., 2015
- New York Times (NYT). Durrett et al., 2016

Implementation:
- TF-IDF + cosine similarity for exemplar retrieval.
- LSTM encoder/decoder.
- Comparable implementation and tuning.
Rouge scores on Gigaword test set

<table>
<thead>
<tr>
<th>Rouge Type</th>
<th>Seq2seq</th>
<th>AttExp, Cao</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td>35.0</td>
<td>36.0</td>
</tr>
<tr>
<td>ROUGE-2</td>
<td>32.4</td>
<td>33.2</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>16.6</td>
<td>17.1</td>
</tr>
</tbody>
</table>

Cao et al., 2018
Rouge scores on Gigaword test set

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-L</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq</td>
<td>35.0</td>
<td>32.4</td>
<td>16.6</td>
</tr>
<tr>
<td>AttExp, Cao</td>
<td>36.0</td>
<td>33.2</td>
<td>17.1</td>
</tr>
<tr>
<td>AdaDec, this work</td>
<td>37.3</td>
<td>34.7</td>
<td>18.5</td>
</tr>
</tbody>
</table>

Cao et al., 2018
Rouge scores on Gigaword test set

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-L</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq</td>
<td>35.0</td>
<td>32.4</td>
<td>16.6</td>
</tr>
<tr>
<td>AttExp, Cao</td>
<td>36.0</td>
<td>33.2</td>
<td>17.1</td>
</tr>
<tr>
<td>AdaDec, this work</td>
<td>37.3</td>
<td>34.7</td>
<td>18.5</td>
</tr>
<tr>
<td>Cao, Full</td>
<td>37.0</td>
<td>34.5</td>
<td>19.0</td>
</tr>
</tbody>
</table>

- Enc. & Att. Exemplar
- Adaptive Decoding
- Rerank

Cao et al., 2018
Rouge scores on NYT test set

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq</td>
<td>41.9</td>
<td>25.1</td>
</tr>
<tr>
<td>Paulus</td>
<td>42.9</td>
<td>26.0</td>
</tr>
<tr>
<td>AttExp</td>
<td>42.5</td>
<td>25.7</td>
</tr>
<tr>
<td>AdaDec</td>
<td>43.2</td>
<td>26.4</td>
</tr>
</tbody>
</table>

Adaptive Decoding

Enc. & Att. Exemplar

Paulus et al., 2018
Experiments: Data2text

Dataset:
- WikiBio. Lebret et al., 2016

Input

<table>
<thead>
<tr>
<th>Jacques-Louis David</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Born</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Died</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Nationality</strong></td>
</tr>
<tr>
<td><strong>Alma mater</strong></td>
</tr>
<tr>
<td><strong>Known for</strong></td>
</tr>
<tr>
<td><strong>Notable work</strong></td>
</tr>
<tr>
<td><strong>Movement</strong></td>
</tr>
<tr>
<td><strong>Awards</strong></td>
</tr>
</tbody>
</table>

Output

Jacques-Louis David (30 August 1748 – 29 December 1825) was a French painter in the Neoclassical style.
Experiments: Data2text

Dataset:
• WikiBio. Lebret et al., 2016

Implementation:
• TF-IDF + cosine similarity for exemplar retrieval.
• LSTM encoder/decoder. See et al., 2017
• Comparable implementation and tuning.
WikiBio test performance

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-4</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq</td>
<td>42.5</td>
<td></td>
</tr>
<tr>
<td>Wiseman</td>
<td>38.6</td>
<td>34.8</td>
</tr>
<tr>
<td>AttExp</td>
<td>40.0</td>
<td>43.1</td>
</tr>
<tr>
<td>AdaDec, this work</td>
<td>40.6</td>
<td>43.6</td>
</tr>
</tbody>
</table>

Wiseman et al., 2018
Conclusion

\[ P(y \mid x, z) \]
Conclusion

Problem

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

Method

\[ \text{AdaDec}_z \left( y \mid \text{Enc}(x) \right) \]

\[ W = \sigma_1 W_1 + \sigma_2 W_2 + \sigma_3 W_3 \]

\[ \begin{align*}
\begin{array}{c}
\text{Rank} \leq d \\
\text{Rank} = 1
\end{array}
\end{align*} \]

\[ d \]
Conclusion

Problem

\[ \mathbb{P}(y \mid x, z) \]

Method

\[
\text{AdaDec}_z \left( y \mid \text{Enc}(x) \right)
\]

\[
W = \sigma_1 W_1 + \sigma_2 W_2 + \sigma_3 W_3
\]

\[
\sigma_1 \quad \text{Rank} \leq d
\]

\[
\sigma_2 \quad \text{Rank} = 1
\]

\[ d \]

Results

<table>
<thead>
<tr>
<th>Rouge</th>
<th>35.0</th>
<th>32.4</th>
<th>36.0</th>
<th>33.2</th>
<th>37.3</th>
<th>34.7</th>
<th>37.0</th>
<th>34.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>39.3</td>
<td>42.5</td>
<td>38.6</td>
<td>34.8</td>
<td>40.0</td>
<td>43.1</td>
<td>40.6</td>
<td>43.6</td>
</tr>
</tbody>
</table>
A Portuguese train derailed in Oporto on Wednesday, killing three people.

<table>
<thead>
<tr>
<th>Exemplars</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
<td>Portuguese train derailed, killing three.</td>
</tr>
<tr>
<td>Two die in a Britain train collision</td>
<td>Three killed in a Portuguese train derailment</td>
</tr>
<tr>
<td>Two people were killed in a Britain train collision</td>
<td>Three people were killed in a Portuguese train derailment</td>
</tr>
<tr>
<td>A train collision in Canada killed two people</td>
<td>A Portugueseses train derails in northern Mexico, killed three</td>
</tr>
</tbody>
</table>