Get To The Point: 
Summarization with Pointer-Generator Networks (ACL 2017)

Authors: Abigail See, Peter J. Liu, Christopher D. Manning 
Presenter: Lu Wang 

[Slides modified from paper conference presentation https://www.aclweb.org/anthology/P17-1099/]
Two approaches to summarization

**Extractive Summarization**

*Select parts (typically sentences) of the original text to form a summary.*

- Easier
- Too restrictive (no paraphrasing)
- Most past work is extractive

**Abstractive Summarization**

*Generate novel sentences using natural language generation techniques.*

- More difficult
- More flexible and human
- Necessary for future progress
Sequence-to-sequence + attention model
Sequence-to-sequence + attention model
Sequence-to-sequence + attention model
Two problems

**Problem 1:** The summaries sometimes reproduce factual details inaccurately.

  e.g. Germany beat Argentina 3-2

  Incorrect rare or out-of-vocabulary word

**Problem 2:** The summaries sometimes repeat themselves.

  e.g. Germany beat Germany beat Germany beat...
Two problems

**Problem 1:** The summaries sometimes reproduce factual details inaccurately.

* e.g. *Germany beat Argentina* 3-2

**Solution:** Use a *pointer* to copy words.

**Problem 2:** The summaries sometimes repeat themselves.

* e.g. *Germany beat Germany beat Germany beat*...
Use pointers

Source Text

Best of both worlds: extraction + abstraction

Pointer-generator network

\[ e'_i = v^T \tanh(W_h h_i + W_s s_i + b_{attn}) \]
\[ d'_i = \text{softmax}(e'_i) \]

Attenions

Final Distribution

Context Vector

Encoder Hidden States

Decoder Hidden States

Vocabulary Distribution

Source Text

Partial Summary
Pointer-generator network
Pointer-generator network

\[ P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} d_i \]
## Improvements

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UNK UNK</strong> was expelled from the dubai open chess tournament</td>
<td><strong>gaioz nigalidze</strong> was expelled from the dubai open chess tournament</td>
</tr>
<tr>
<td>the 2015 rio olympic games</td>
<td>the 2016 rio olympic games</td>
</tr>
</tbody>
</table>
Two problems

**Problem 1:** The summaries sometimes reproduce factual details inaccurately.

  e.g. *Germany beat Argentina 3-2*

**Solution:** Use a pointer to copy words.

**Problem 2:** The summaries sometimes repeat themselves.

  e.g. *Germany beat Germany beat Germany beat...*

**Solution:** Penalize repeatedly attending to same parts of the source text.
Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday
Summary: Germany beat _____

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Don’t attend here

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday
Summary: Germany beat _____

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

Reducing repetition with coverage

**Coverage** = cumulative attention = what has been covered so far

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday
Summary: Germany beat _____

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

---

Reducing repetition with coverage

Coverage = cumulative attention = what has been covered so far

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday
Summary: Germany beat _____

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

\[ c_t^t = \sum_{t'=0}^{t-1} d_t^t \]

\[ \text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(d_i^t, c_i^t) \]

\[ \text{covloss}_t = \sum_i \min(d_i^t, c_i^t) \]

Datasets

• CNN/Daily Mail (Hermann et al., 2015)
  • 287,226 training examples, 13,368 validation examples and 11,490 testing examples
  • limit the input length to 400 tokens and output length to 100 tokens for training and 120 for testing
Results

ROUGE compares the machine-generated summary to the human-written reference summary and counts co-occurrence of 1-grams, 2-grams, and longest common sequence.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nallapati et al. 2016</td>
<td>35.5</td>
<td>13.3</td>
<td>32.7</td>
</tr>
<tr>
<td>Ours (seq2seq baseline)</td>
<td>31.3</td>
<td>11.8</td>
<td>28.8</td>
</tr>
<tr>
<td>Ours (pointer-generator)</td>
<td>36.4</td>
<td>15.7</td>
<td>33.4</td>
</tr>
<tr>
<td>Ours (pointer-generator + coverage)</td>
<td><strong>39.5</strong></td>
<td><strong>17.3</strong></td>
<td><strong>36.4</strong></td>
</tr>
</tbody>
</table>

Previous best abstractive result

Our improvements
Results

ROUGE compares the **machine-generated summary** to the **human-written reference summary** and counts co-occurrence of **1-grams**, **2-grams**, and **longest common sequence**.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nallapati et al. 2016</td>
<td>35.5</td>
<td>13.3</td>
<td>32.7</td>
</tr>
<tr>
<td>Ours (seq2seq baseline)</td>
<td>31.3</td>
<td>11.8</td>
<td>28.8</td>
</tr>
<tr>
<td>Ours (pointer-generator)</td>
<td>36.4</td>
<td>15.7</td>
<td>33.4</td>
</tr>
<tr>
<td>Ours (pointer-generator + coverage)</td>
<td><strong>39.5</strong></td>
<td><strong>17.3</strong></td>
<td><strong>36.4</strong></td>
</tr>
</tbody>
</table>

Lead-3 (first three sentences) 40.3 17.7 36.6

Previous best abstractive result

Our improvements

The difficulty of evaluating summarization

• Summarization is subjective
  • There are many correct ways to summarize

• ROUGE is based on strict comparison to a reference summary
  • Intolerant to rephrasing
  • Rewards extractive strategies

• Take first 3 sentences as summary → higher ROUGE than (almost) any published system
  • Partially due to news article structure
A Deep Reinforced Model for Abstractive Summarization (ICLR 2018)

Authors: Romain Paulus, Caiming Xiong, Richard Socher
Presenter: Lu Wang

[Some figures taken from Paulus’ presentation]
Three problems

• Repetitive content in the output (this is discussed in the first paper)
Three problems

• Repetitive content in the output (this is discussed in the first paper)

• Long-term coherence
  • hard to stay on the same topic or show connections when multiple sentences are generated
  • Ordering 1: Lisa went to sail. There was a gale. Lisa came home.
  • Ordering 2: Lisa came home. There was a gale. Lisa went to sail.
Three problems

• Repetitive content in the output (this is discussed in the first paper)

• Long-term coherence
  • hard to stay on the same topic or show connections when multiple sentences are generated
  • Ordering 1: Lisa went to sail. There was a gale. Lisa came home.
  • Ordering 2: Lisa came home. There was a gale. Lisa went to sail.

• Directly optimize on ROUGE scores
  • ROUGE measure word overlaps between system generated summaries and human-written summaries
  • existing training objective use likelihood of each generated token, i.e. $p(y_t|x)$
Three problems

• Repetitive content in the output

• Long-term coherence

• Directly optimize on ROUGE scores
Temporal attention on the input + self-attention on the output
Temporal attention

• Input attention weights (different from pointer-generator paper)

\[ e_{ti} = f(h_t^d, h_i^e) \]

\[ f(h_t^d, h_i^e) = h_t^d W_{\text{attn}} h_i^e. \]

• Penalizing the tokens that obtained high attentions

\[ e'_{ti} = \begin{cases} 
\exp(e_{ti})/\sum_{j=1}^{t-1} \exp(e_{ji}) & \text{if } t = 1 \\
\exp(e_{ti})/\sum_{j=1}^{t-1} \exp(e_{ji}) & \text{otherwise}
\end{cases} \]
Self-attention (intra-decoder attention)

• How to be aware of what has been generated?
Self-attention (intra-decoder attention)

Consider what has been generated

\[
c_t^d = \sum_{j=1}^{t-1} \alpha_{tj}^d h_j^d
\]

\[
e_{tt'}^d = h_t^d W_{\text{attn}}^d h_{t'}^d
\]

\[
\alpha_{tt'}^d = \frac{\exp(e_{tt'}^d)}{\sum_{j=1}^{t-1} \exp(e_{tj}^d)}
\]
Self-attention (intra-decoder attention)

\[
e_{tt'}^d = h_t^d h_t'^T W_{\text{attn}} h_t^d
\]
\[
\alpha_{tt'}^d = \frac{\exp(e_{tt'}^d)}{\sum_{j=1}^{t-1} \exp(e_{tj}^d)}
\]
\[
c_t^d = \sum_{j=1}^{t-1} \alpha_{tj}^d h_j^d
\]

\[
p(y_t \mid u_t = 0) = \text{softmax}(W_{\text{out}}[h_t^d \| c_t^e \| c_t^d] + b_{\text{out}})
\]

Input attention  self-attention

Encoder  Decoder
Three problems

• Repetitive content in the output

• Long-term coherence

• Directly optimize on ROUGE scores
Global reward with ROUGE

- Idea: directly using ROUGE scores as reward
- But ROUGE is not differentiable

- Training method: self-critical **policy gradient** training algorithm

\[
L_{rl} = (r(y) - r(y^s)) \sum_{t=1}^{n'} \log p(y^s_t | y^s_1, \ldots, y^s_{t-1}, x)
\]

Baseline: greedy decoding  
Sampled sequence

- Intuitively, we aim to maximize the conditional likelihood of the **sampled sequence** \(y^s\) if it obtains a higher reward than the **baseline**
New training objective

\[ L_{mixed} = \gamma L_{rl} + (1 - \gamma) L_{ml} \]

New reinforcement learning objective  Regular log-likelihood objective
Datasets for experiments

• CNN/Daily Mail (Hermann et al., 2015)
  • 287,113 training examples, 13,368 validation examples and 11,490 testing examples
  • limit the input length to 800 tokens and output length to 100 tokens

• New York Times (Sandhaus, 2008)
  • 589,284 examples for training, 32,736 for validation, and 32,739 for testing
Results on CNN/Daily Mail

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead-3 (Nallapati et al., 2017)</td>
<td>39.2</td>
<td>15.7</td>
<td>35.5</td>
</tr>
<tr>
<td>SummaRuNNer (Nallapati et al., 2017)</td>
<td>39.6</td>
<td>16.2</td>
<td>35.3</td>
</tr>
<tr>
<td>words-lvt2k-temp-att (Nallapati et al., 2016)</td>
<td>35.46</td>
<td>13.30</td>
<td>32.65</td>
</tr>
<tr>
<td>ML, no intra-attention</td>
<td>37.86</td>
<td>14.69</td>
<td>34.99</td>
</tr>
<tr>
<td>ML, with intra-attention</td>
<td>38.30</td>
<td>14.81</td>
<td>35.49</td>
</tr>
<tr>
<td>RL, with intra-attention</td>
<td>41.16</td>
<td>15.75</td>
<td>39.08</td>
</tr>
<tr>
<td>ML+RL, with intra-attention</td>
<td>39.87</td>
<td>15.82</td>
<td>36.90</td>
</tr>
</tbody>
</table>
ROUGE-1 improvement by adding intra-attention on CNN/Daily Mail
## Results on NYT

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML, no intra-attention</td>
<td>44.26</td>
<td>27.43</td>
<td>40.41</td>
</tr>
<tr>
<td>ML, with intra-attention</td>
<td>43.86</td>
<td>27.10</td>
<td>40.11</td>
</tr>
<tr>
<td>RL, no intra-attention</td>
<td>47.22</td>
<td>30.51</td>
<td>43.27</td>
</tr>
<tr>
<td>ML+RL, no intra-attention</td>
<td>47.03</td>
<td>30.72</td>
<td>43.10</td>
</tr>
</tbody>
</table>
**Ground truth summary**

Button denied 100th race start for McLaren after ERS failure. Button then spent much of the Bahrain Grand Prix on Twitter delivering his verdict on the action as it unfolded. Lewis Hamilton has out-qualified and finished ahead of Mercedes team-mate Nico Rosberg at every race this season. Bernie Ecclestone confirms F1 will make its bow in Azerbaijan next season.

**ML, with intra-attention (ROUGE-1 41.58)**

Button was denied his 100th race for McLaren. ERS prevented him from making it to the start-line. The Briton. He quizzed after Nico Rosberg accused Lewis Hamilton of pulling off such a manoeuvre in China. Button has been in Azerbaijan for the first time since 2013.

**RL, with intra-attention (ROUGE-1 50.00)**

Button was denied his 100th race for McLaren after an ERS prevented him from making it to the start-line. It capped a miserable weekend for the Briton. Button has out-qualified. Finished ahead of Nico Rosberg at Bahrain. Lewis Hamilton has. In 11 races. The race. To lead 2,000 laps. In. . . And. . .

**ML+RL, with intra-attention (ROUGE-1 44.00)**

Button was denied his 100th race for McLaren. The ERS prevented him from making it to the start-line. Button was his team mate in the 11 races in Bahrain. He quizzed after Nico Rosberg accused Lewis Hamilton of pulling off such a manoeuvre in China.
Human evaluation on CNN/Daily Mail

<table>
<thead>
<tr>
<th>Model</th>
<th>Readability</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>6.76</td>
<td>7.14</td>
</tr>
<tr>
<td>RL</td>
<td>4.18</td>
<td>6.32</td>
</tr>
<tr>
<td>ML+RL</td>
<td>7.04</td>
<td>7.45</td>
</tr>
</tbody>
</table>
Conclusion

• Intra-decoder attention helps with long summary generation

• Reinforcement learning with ROUGE as reward improves performance

• Simply using reinforcement learning hurts readability