Incorporating Topic Sentence on Neural News Headline Generation

Jan Wira Gotama Putra¹, Hayato Kobayashi²,³, Nobuyuki Shimizu²

¹Tokyo Institute of Technology  ²Yahoo Japan Corporation  ³RIKEN AIP

gotama.w.aa@m.titech.ac.jp  {hakobaya, nobushim}@yahoo-corp.jp

*) Research was done when the first author was an intern (summer 2017) at Yahoo! Japan
Automatic Headline Generation

- Given a news document, we want to generate a corresponding headline

- Automatic headline generation system is used by news editor as a supporting tool

- Single document summarization
  - Extractive approach (Zajic et al., 2004; Colmenares et al., 2015)
  - Abstractive approach (Banko, et al., 2000; Rush et al. 2015)

Incorporating Topic Sentence on Neural News Headline Generation

https://www.japantimes.co.jp/life/2018/03/04/lifestyle/traditional-arts-live-kids/#.WqFfBZOuxsM
Abstractive Headline Generation

• Abstractive approach recently motivated by the success of neural machine translation systems (sequence to sequence) (Sutskever et al., 2014)

• Formalization
  • Given a sequence of $N$ input words (source documents)
    $$x = x_1, x_2, ..., x_N$$
  • The task is to find a sequence of $M$ output words (summary/headline)
    $$y = y_1, y_2, ..., y_M; M < N$$
  • It means we are modeling the conditional probability of input—output pair
    $$\text{summary} = \arg\max_y P(y|x, \theta)$$
Factoring the Objective

\[
P(y|x, \theta) = \prod_{t=1}^{M} P(y_t | \{y_1, \ldots, y_t\}, x, \theta)
\]

A decoder converts the representation of input (c) into a sequence of output y

Encoder converts a sequence of input x into a single representation c
Encoder – Decoder Model
Encoder – Decoder Model

Input Representation \( c \)

Backward RNN

\( x_1 \)

\( \ldots \)

\( x_{N-1} \)

\( x_N \)

Forward RNN

Incorporating Topic Sentence on Neural News Headline Generation
Encoder – Decoder Model

Training objective: minimalizing loss

\[ L = - \sum_{(x,y) \in Dataset} P(y \mid x, \theta) \]
Encoder – Decoder Model with Attention

Input Representation $\mathbf{c}$

Attention

Forward RNN

$x_1$ $x_{N-1}$ $x_N$

Backward RNN

$y_1$ $y_2$ $y_{M-1}$ $y_M$

$\langle$EOS$\rangle$

Training objective: minimalizing loss

$$L = - \sum_{(x,y) \in \text{Dataset}} P(y | x, \theta)$$
Related Work

Long Input
Vanishing gradient problem
(Cho et al., 2014; Tan et al., 2017)

Past Studies (headline generation)
Use the first sentence
(Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; Ayana et al., 2017)
Problems

• The first sentence might not be effective, as the information in a text is distributed across sentences (Alfonseca et al., 2013)

• Using long input may degrade the performance of encoder-decoder (Cho et al., 2014; Tan et al., 2017)

• Previous studies did not consider 5W1H (what, who, when, where, whom, how) information when analyzing news (Wang, 2012).

• How to consider inverse pyramid structure of news (organization structure)
Proposal (contribution)

• Using *topic sentence* instead of/in addition to the first sentence

• *Topic sentence* (Wang, 2012) contains *key information of news*; it has the *<subject, verb, object>* elements and at least one subordinate element *time* or *location* (factual information).
  
  • *Time = DATE and TIME* (NE tag)
  • *Location = GPE and LOC* (NE tag)

• We extract only one topic sentence from news (the earliest sentence satisfying the rules)
Proposal (contribution)

Past Studies (headline generation)
Use the first sentence (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; Ayana et al., 2017)

Current Study
Use topic sentence for sentence selection
Hypothesis

• We hypothesized that topic sentence is likely to provide a better generalization for the encoder–decoder than using the first sentence.

• Generalization means allowing the model to predict the headline of the unseen data in a better way.

• Topic sentence ≠ statistical ranking techniques (SRT); SRT considers surface information without considering factual information.
1. Is the topic sentence more useful than the first sentence for headline generation?

2. Is the topic sentence helpful in addition to the first sentence for headline generation?
Experimental Setting

• We train the encoder—decoder model using three variants of input
  • First sentence (OF)
  • Topic sentence (OT)
  • Both first and topic sentence (OTF)

• We extract only one topic sentence (the earliest sentence satisfying the rules)

• We use the seq2seq implementation of OpenNMT (Klein, et al; 2017)
  • Encoder is 2-layer bidirectional LSTM RNN (500 hidden units)
  • Decoder is 2-layer LSTM RNN (500 hidden units)
  • Global attention mechanism and dropout (0.3) are used
Dataset

• We used Gigaword dataset (10M documents)

<table>
<thead>
<tr>
<th>Data</th>
<th># docs</th>
<th>Found-1</th>
<th>Found-2-</th>
<th>Not found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train (~90%)</td>
<td>2,755K</td>
<td>2,023K (73.43%)</td>
<td>580K (21.06%)</td>
<td>152K (5.54%)</td>
</tr>
<tr>
<td>Valid (~5%)</td>
<td>139K</td>
<td>101K (72.76%)</td>
<td>29K (21.58%)</td>
<td>7K (5.69%)</td>
</tr>
<tr>
<td>Test (~5%)</td>
<td>134K</td>
<td>98K (72.91%)</td>
<td>28K (21.19%)</td>
<td>8K (5.90%)</td>
</tr>
</tbody>
</table>

• Found-1 : Topic sentence is found as the first sentence of the text
• Found-2  : Topic sentence is found as the second or later sentence of the text
• Not found: Topic sentence is not found in the text
## Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set</th>
<th>Topic</th>
<th>First</th>
<th>First and Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-I</td>
<td>R-2</td>
<td>R-L</td>
<td>Copy rate</td>
</tr>
<tr>
<td>OF</td>
<td>29.45</td>
<td>12.06</td>
<td>26.97</td>
<td>0.72</td>
</tr>
<tr>
<td>OT</td>
<td>33.73</td>
<td>14.37</td>
<td>30.77</td>
<td>0.71</td>
</tr>
<tr>
<td>OTF</td>
<td>32.00</td>
<td>13.03</td>
<td>29.11</td>
<td>0.76</td>
</tr>
</tbody>
</table>

- **OF**: trained using (first sentence – headline)
- **OT**: trained using (topic sentence – headline)
- **OTF**: trained using (both topic+ first sentences – headline pair)
- **R**: ROUGE
Output Example

• **Input:** for american consumers, the prospect of falling prices sure sounds like a good thing but a prolonged and widespread decline, with everything from real-estate values to income collapsing, would spell disaster for the u.s. economy.

• **Reference headline:** falling prices stagnant employment numbers have economists worrying about deflation

• **OF Prediction:** u.s. consumer confidence drops to new high

• **OT Prediction:** u.s. consumer prices fall ## percent in may

• **OTF Prediction:** u.s. consumer prices fall for first time since ####
### Additional Test

<table>
<thead>
<tr>
<th>Model</th>
<th>Training data</th>
<th>ROUGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OF</td>
<td>2.7 M docs (Rush et al., 2015 + additional filter)</td>
<td></td>
</tr>
<tr>
<td>OT</td>
<td>2.7 M docs (Rush et al., 2015 + additional filter)</td>
<td></td>
</tr>
<tr>
<td>OTF</td>
<td>2.7 M docs (Rush et al., 2015 + additional filter)</td>
<td></td>
</tr>
<tr>
<td>ABS+</td>
<td>3.7 M docs (Rush et al., 2015)</td>
<td></td>
</tr>
<tr>
<td>words-lvt2k-1sent</td>
<td>3.7 M docs (Rush et al., 2015)</td>
<td></td>
</tr>
<tr>
<td>OpenNMT bechmark*</td>
<td>3.7 M docs (Rush et al., 2015)</td>
<td></td>
</tr>
<tr>
<td>RAS-Elman</td>
<td>3.7 M docs (Rush et al., 2015)</td>
<td></td>
</tr>
<tr>
<td>MRT</td>
<td>3.7 M docs (Rush et al., 2015)</td>
<td></td>
</tr>
</tbody>
</table>

#### Small Test Set

Conclusion

1. Is the topic sentence more useful than the first sentence for headline generation?
   Yes, for training (generalization)

2. Is the topic sentence helpful in addition to the first sentence for headline generation?
   Yes, it acts as a supporting device
Future Direction

1. Assess the difference of using topic sentence as opposed to other sentence selection/ranking methods

2. Investigate whether using/adding other types of subset of the full news document is able to improve the performance

3. Automatically decide the optimal subset of text as input for headline generation (encoder-decoder architecture)


References (2)


