A Case Study on Neural Headline Generation for Editing Support

Kazuma Murao*,1, Ken Kobayashi*,1, Hayato Kobayashi1,2, Taichi Yatsuka1, Takeshi Masuyama1, Tatsuru Higurashi1, Yoshimune Tabuchi1

1Yahoo Japan Corporation 2RIKEN AIP (*Equal contribution)
Summary

• Our work
  • Address “short title” generation for a news aggregation service, where editors create short titles to introduce important articles

• Contributions
  • Show a practical use case of neural headline generation
    • Most news articles basically already have headlines
  • Propose an encoder-decoder model with multiple encoders
  • Deploy our model to an editing support tool and show the results of comparing the editors’ behavior
Yahoo! News

- Biggest news portal in Japan
- PV/month: 15,000,000,000+
- Editors’ choice feature -

1. Pick up important news articles
2. Put a new shorter headline, called short title

Pros:
- Quick understandability
- Saving display space
Short title generation as editing support

- **Purpose**: To generate short title candidates to help editors
- **Task**: Translation from (headline, lead) to short title
  - Lead is a short version (summary) of the article

---

**Selected news article**

**Headline**
2016年「生理学・医学賞」は誰の手に？
日本科学未来館がノーベル賞予想

**Lead**
2016年のノーベル賞発表まで一週間を切りました。10月3日の生理学・医学賞を皮切りに、4日には物理学賞、5日には化学賞、6日には生物学賞、7日には化学賞を発表します。

**Short title**
ノーベル賞 今年は誰の手に？

**List of news articles**

4人死傷 容疑者は元千葉市議
富士フイルム 賠償請求も検討

---

4/19

Copyright © 2019 Yahoo Japan Corporation. All Rights Reserved.
### Example of (short title, headline, lead)

<table>
<thead>
<tr>
<th>Japanese</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>首相「忖度ないと言い切れず」</td>
<td>The prime minister cannot say that there is no surmise.</td>
</tr>
<tr>
<td>村田ら「忖度なかったと言い切ることはできない＝加計問題で安倍首相」</td>
<td>It cannot be said that there is no “sontaku” with absolute certainty. The prime minister Abe said about the problem of “Kake Gakuen (Kake school)”</td>
</tr>
</tbody>
</table>

Prime Minister Shinzo Abe said, in an intensive deliberation with the House of Councilors Budget Committee held on the afternoon of the 14th, as an answer to a question about whether bureaucrats surmised to the prime minister regarding the Kake suspicion, “It is difficult to understand whether there is a sontaku (surmise)” . He said “It cannot be said that there was nothing wrong,” while explaining that “I do not need to be obsequious”. An answer to Ichiro Tsukada (LDP).

---

**Short title generation task is not so easy**

*Lengths are different*

*Phrase order is changed*
Encoder-decoder model with attention

- Conditional language model consisting of two RNNs
- Described by three components (encoder, attention, decoder)

Attention calculates a context $c_t$ from the encoder’s states $h_s$.

$$c_t = \sum_{s=1}^{S} a_t(s) h_s$$

$$p(y_{t+1} \mid y_{\leq t}, x) = g_{\text{dec}}(\hat{h}_t, c_t)$$
Proposed method: GateFusion

- Combine headline and lead contexts w/ gating mechanism

Headline Enc. → Atten. \(d_t\) → Gate \(\bar{c}_t\) → Decoder

Lead Enc. → Atten. \(d'_t\)

- Existing work (Hori+ 2017) used an attention mechanism
  \[ \bar{c}_t = \alpha d_t + \beta d'_t \]

- Gating mechanism:
  \[ w_t = \sigma(W[d_t; d'_t; \hat{h}_t]), \]
  \[ w'_t = \sigma(W'[d_t; d'_t; \hat{h}_t]), \]
  \[ \bar{c}_t = w_t \odot d_t + w'_t \odot d'_t, \]

- Fusion based on scalar weights

- Fusion based on vector weights
Baselines with multiple encoders

- Multi-modal method (Hori+ 2017)
  - Headline Enc. → Atten. → $d_t$ → Atten. → $\tilde{c}_t$ → Decoder
  - Lead Enc. → Atten. → $d_t'$ → Atten. → $\tilde{c}_t$ → Decoder

- Query-based method (Nema+ 2017)
  - Lead Enc. → Atten. → $d_t'$ (query) → Atten. → $\tilde{c}_t$ → Decoder
  - (main source) Headline Enc. → Atten. → Adjust weights → $\tilde{c}_t$ → Decoder

Fusion based on scalar weights
Fusion based on cascade connection
Adjust weights
Training dataset

• 263K triples of (headline, lead, short title) in Yahoo! News
  • Training (90%), validation (5%), testing (5%)

• Statistics:

<table>
<thead>
<tr>
<th></th>
<th>Headline</th>
<th>Lead</th>
<th>Short title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average length</td>
<td>24.87</td>
<td>128.49</td>
<td>13.05</td>
</tr>
<tr>
<td>Character type</td>
<td>3618</td>
<td>4226</td>
<td>3156</td>
</tr>
</tbody>
</table>

• Extractively solvable instances: 20%
  • Characters in each short title are completely covered by the headline
• Edit distance of headlines and short titles: 23.74
  • Short titles cannot be easily created only from headlines
Model and training settings

- Implemented on OpenNMT
- Headline encoder: BiLSTM
- Lead encoder: CNN (Kim, 2014)
  - To reduce the computational time
- Ensemble of 10 models
- Hyper-parameter settings are listed in the right table

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of layers (RNN, CNN)</td>
<td>3</td>
</tr>
<tr>
<td># of units (embedding)</td>
<td>200</td>
</tr>
<tr>
<td># of units (RNN, CNN)</td>
<td>400</td>
</tr>
<tr>
<td># of units (context)</td>
<td>400</td>
</tr>
<tr>
<td>Window width of CNN</td>
<td>7</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>0.3</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.05</td>
</tr>
<tr>
<td>Momentum rate</td>
<td>0.8</td>
</tr>
<tr>
<td>Learning decay rate</td>
<td>0.85</td>
</tr>
<tr>
<td># of epochs</td>
<td>20</td>
</tr>
<tr>
<td>Batch size</td>
<td>64</td>
</tr>
<tr>
<td>Beam width</td>
<td>5</td>
</tr>
</tbody>
</table>
Human evaluation by crowdsourcing

- Two crowdsourcing tasks for readability and usefulness
  - Average score of 10 workers for each of 1,000 outputs

- Readability (four-point scale)
  - How readable a short title was

- Usefulness (four-point scale)
  - How useful a short title was compared to the headline
Evaluation results (1/2)

- Our model performed well for the usefulness measure

<table>
<thead>
<tr>
<th></th>
<th>Readability</th>
<th>Usefulness</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor</td>
<td>3.62</td>
<td>3.18</td>
<td>3.40</td>
</tr>
<tr>
<td>Prefix</td>
<td>2.72</td>
<td>2.38</td>
<td>2.55</td>
</tr>
<tr>
<td>OpenNMT</td>
<td>3.53</td>
<td>3.16</td>
<td>3.35</td>
</tr>
<tr>
<td>MultiModal</td>
<td>3.51</td>
<td>3.12</td>
<td>3.32</td>
</tr>
<tr>
<td>QueryBased</td>
<td>3.52</td>
<td>3.11</td>
<td>3.32</td>
</tr>
<tr>
<td><strong>GateFusion</strong></td>
<td><strong>3.50</strong></td>
<td><strong>3.22</strong></td>
<td><strong>3.36</strong></td>
</tr>
</tbody>
</table>

\[ \text{Average} = \frac{\text{Readability} + \text{Usefulness}}{2} \]

- Complicated expressions
- Aggressively copy characters
Evaluation results (2/2)

- Our model performed well for the usefulness measure

<table>
<thead>
<tr>
<th></th>
<th>Readability</th>
<th>Usefulness</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor</td>
<td>3.62</td>
<td>3.18</td>
<td>3.40</td>
</tr>
<tr>
<td>Prefix</td>
<td>2.72</td>
<td>2.38</td>
<td>2.55</td>
</tr>
<tr>
<td>OpenNMT</td>
<td>3.53</td>
<td>3.16</td>
<td>3.35</td>
</tr>
<tr>
<td>MultiModal</td>
<td>3.51</td>
<td>3.12</td>
<td>3.32</td>
</tr>
<tr>
<td>QueryBased</td>
<td>3.52</td>
<td>3.11</td>
<td>3.32</td>
</tr>
<tr>
<td>GateFusion</td>
<td>3.50</td>
<td>3.22</td>
<td>3.36</td>
</tr>
<tr>
<td>HybridFusion</td>
<td>3.55</td>
<td>3.22</td>
<td>3.39</td>
</tr>
</tbody>
</table>

Correct titles
First 13 chars
Single enc.

Multi enc.
Our models
Gate+Query

Close to Editor
QueryBased helped GateFusion generate headline-style outputs

Close to Editor
| Input and generated title (Japanese) | Evolution of Darvish: turning adversity into opportunity.  
|-------------------------------------|------------------------------------------------------|
| **Headline**                       | Evolution of Darvish, turning adversity into opportunity.  
| **Lead**                           | Yu Darvish (29) in Rangers took a mound for the first time in 1 year and 9 months with Pirates [...]  
| **Editor**                         | Dar sculpted his body better than before surgery.  
| **OpenNMT**                        | Evolution, turning adversity into opportunity.  
| **HybridFusion**                   | Dar turned adversity into opportunity.  

Our model worked even in this real-world application.
Editing support tool

• Editors can check candidates when creating short titles

1. Enter a URL
2. Fetch the content
3. Display candidates
4. Copy & edit a candidate
Functionalities in the tool

• Cutoff unpromising candidates
  • If perplexity > x
  • To keep the system quality
• Skipping redundant candidates
  • If edit distance < y
  • To display various outputs
• Highlighting unknown characters
  • If not in the article
  • To encourage fact checking
Effect of the tool release

• Editors’ behavior in three weeks before/after the release
• Rate at which an editor’s title matches the generated one by X%

Editors began to refer to generated outputs after the release

**Rate of 100% match titles**

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>3.8%</td>
<td>6.1%</td>
</tr>
</tbody>
</table>

Before: 3.8% → After: 6.1% (x1.6)

**Rate of 80+% match titles**

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>14.0%</td>
<td>18.5%</td>
</tr>
</tbody>
</table>

Before: 14.0% → After: 18.5% (x1.3)
Conclusion

• Short titles were successfully generated for editing support
• Editors began to refer to generated titles of our system

• Future work
  • Verify how much our model can affect click-through rate
    • Need a much safer model to avoid generating fake titles

• Acknowledgements
  • We would like to thank editors and engineers in the news service who continuously supported our experiments
Thank you for your attention!