Neural Headline Generation with Self-Training

Kazuma Murao, Shintaro Takemae, Hayato Kobayashi, Taichi Yatsuka, Masaki Noguchi, Hitoshi Nishikawa, Takenobu Tokunaga
1: Yahoo Japan Corporation, 2: Tokyo Institute of Technology, 3: Riken AIP
kmurao@yahoo-corp.jp

Neural Headline Generation

Our challenge in news-aggregator in Yahoo! JAPAN
Condensed headline is used to be shown in the Yahoo! JAPAN website as a hyperlink to a news article instead of the original headline and to allow its visitors to swiftly determine if they should read the article. Editing assistant system which shows candidates of condensed headline to editors would be useful.

Condensed headline generation with encoder-decoder model
Recent automatic summarization systems depend heavily on machine learning which based on a lot of training examples. However, since the condensed headlines are manually written by human workers in addition to their original headlines, making them is costly, and therefore it is not easy to prepare them in large quantities.

<table>
<thead>
<tr>
<th>Original Headline</th>
<th>Condensed Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;天皇陛下&gt;退位後に「赤坂御用地」に転居 - 安居内宮署旗</td>
<td>陛下 退位後に赤坂に転居</td>
</tr>
</tbody>
</table>

Proposed Method

Step 1. Train baseline model
A baseline model is learned from pairs of headlines and their corresponding condensed headlines.

Step 2. Generate pseudo headlines
We run the baseline model to generate pseudo condensed headlines from the headlines without condensed ones.

Step 3. Train proposed model
Our proposed model is learned from both original and pseudo condensed headlines.

Experiments

◆ Implementation and settings
  • Generate condensed headline from original one
  • Baseline method employs basic encoder–decoder model with an attention mechanism (with LSTM, character based). Developed with OpenNMT-Toolkit.

◆ Data
  • Original and condensed headlines: about 100K
    - Split into: Training 60%, Validation 20%, Test 20%
  • Original headlines: 2M
  • Compare number of unlabeled training examples from 200K to 2M
    * Gathered by web crawling

◆ Compared method
  • Baseline method
  • Pre-training method [Dai and Le, NIPS 2015]
  • Proposed method (with 600K unlabeled headlines)

◆ Results
  Proposed method significantly overcome baseline method. (p<0.05)

◆ Discussion
  Too much unlabeled training examples generated from an initial model overweights the labeled examples (right figure). This implies that if we could filter out the inferior training examples and choose only good training examples, we could improve the accuracy.
  Consideration of news description
  Human editors carefully read the body of articles as well as their headlines in order to make their condensed headlines, and therefore information written in the body should also be considered to generate better condensed headlines.

Results

<table>
<thead>
<tr>
<th>Original Headline</th>
<th>Gold Standard</th>
<th>Baseline (ROUGE-1)</th>
<th>Proposed (ROUGE-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>日本のロボット産業、なぜ急速に進化したか</td>
<td>ソフトバンクがロボット事業に参入</td>
<td>日本のロボット産業、なぜ急速に進化したか</td>
<td></td>
</tr>
</tbody>
</table>

Baseline

Pre-Training

Pre-Training (all weights)

Proposed (600K)

ROUGE-1 result

<table>
<thead>
<tr>
<th>ROUGE-1 result</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.571</td>
</tr>
<tr>
<td>Pre-Training(w/o attention weights)</td>
<td>0.520</td>
</tr>
<tr>
<td>Pre-Training(all weights)</td>
<td>0.503</td>
</tr>
<tr>
<td>Proposed (600K)</td>
<td>0.574</td>
</tr>
</tbody>
</table>