**STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset**

Yuya Yoshikawa, Yutaro Shigeto, Akikazu Takeuchi (STAIR Lab, Chiba Institute of Technology, Japan)

1. **Paper Summary**

- We developed an image caption dataset, STAIR Captions, which is the largest Japanese dataset and has 820,310 Japanese captions for all the MS-COCO images.
- We confirmed that a neural network trained using STAIR Captions can generate more natural and better Japanese captions, compared to those generated using En-Ja MT after generating English captions.

2. **Motivation**

**Why we developed STAIR Captions**

- **Problem: low Japanese resources for image captioning**
  - Most datasets are annotated in English.
  - YJ Captions [Miyazaki+ ACL2016] is a Japanese caption dataset, but they annotated captions for the small part of MS-COCO images.

- **Features:**
  1. Send!
  2. Look at image
  3. Write description

3. **STAIR Captions**

**Guidelines and procedure of annotations, and dataset statistics**

For all the images in 2014 edition of MS-COCO, we annotated Japanese captions by about 2,100 crowdsourcing and part-time job workers in a half year.

**Annotation system.** We developed a web-system for annotation.

1) Look at image

2) Write description

3) Send!

**Annotation guidelines.**

When annotation, we asked the workers to follow our annotation guidelines.

- **Guidelines:**
  1. A caption must contain more than 15 letters.
  2. A caption must follow the da/dearu style (one of writing styles in Japanese).
  3. A caption must describe only what is happening in an image and the things displayed therein.
  4. A caption must be a single sentence.
  5. A caption must not include emotions or opinions about the image.

**Quality control.** For randomly extracted captions (1-2% of the whole captions), we checked whether the captions follow the guidelines. If not, we removed the captions.

**Comparison of dataset statistics.**

Compared to YJ Captions, STAIR Captions has

- 6.19x (4.67x) annotated images
- 6.23x (4.65x) Japanese captions
- 2.69x (2.41x) vocabulary size

*Numbers in brackets denote the sizes in public part.

4. **Experiments**

**Comparing the performance of image captioning in Japanese**

**Configuration.** We compare two methods using the same neural network (NN) architecture.

- **En-generator → MT:** after generating English captions using NN learned on MS-COCO, translates the captions into Japanese ones by Google Translate (GNMT version).
- **Ja-generator:** generates Japanese captions directly by NN learned on STAIR Captions

**Neural network architecture.**

We used **NeuralTalk2** [Karpathy+ 2015]

- **Encoder:** VGG
- **Decoder:** LSTM

**Optimization.** We learned LSTM parameters by mini-batch RMSProp (mini-batch size = 20), while CNN parameters pre-trained on ImageNet are fixed.

**Quantitative result.**

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>ROUGE_L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-generator → MT</td>
<td>0.565</td>
<td>0.330</td>
<td>0.204</td>
<td>0.127</td>
<td>0.449</td>
<td>0.324</td>
</tr>
<tr>
<td>Ja-generator</td>
<td>0.763</td>
<td>0.614</td>
<td>0.492</td>
<td>0.385</td>
<td>0.553</td>
<td>0.833</td>
</tr>
</tbody>
</table>

- Ja-generator outperforms En-generator → MT in terms of all the metrics.
- Future work: comparing this performance with the one using YJ Captions.

**Typical examples.**

- **En-generator → MT:** En-generator can generate natural captions in English, but, after translating the captions into Japanese ones by MT, the captions in often change unnatural ones (because some phrases are translated word-by-word).
- **Ja-generator:** can generate natural phrases and select appropriate vocabularies.

**Input:** image

**Translator (e.g. NN)**

- a white and light gray kitchen with stove, sink, and refrigerator.

**Output:** description (text) from an image.

---

**Note:**)

http://captions.stair.center

**Image:** Available on both PC and smartphones

*Numbers in brackets denote the sizes in public part.*

**Image:** Detect too short captions automatically

**Future work:**

- Comparing this performance with the one using YJ Captions
- Future work: comparing this performance with the one using YJ Captions