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

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

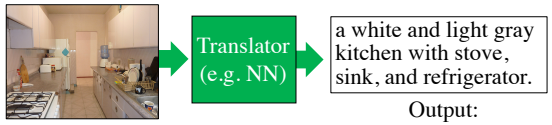
STAIR Captions is available for download!
http://captions.stair.center

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

Image captioning is to automatically generate a description (text) from an image.



Input: image — Output: description (text)

Problem: low Japanese resources for image captioning

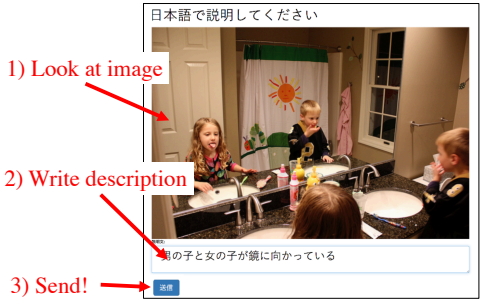
- Most datasets are annotated in English
- **YJ Captions [Miyazaki+ ACL2016]** is a Japanese caption dataset, but they annotated captions for the small part of MS-COCO images
- Q: Why don't you translate English captions into Japanese ones?
  A: MT often generates unnatural translations for captions

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.



Features:

- Available on both PC and smartphones
- Detect too short captions automatically
- Do not display the same images again for a worker

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

	STAIR Captions	YJ Captions
# of images	164,062 (123,287)	26,500
# of captions	820,310 (616,435)	131,740
Vocabulary size	35,642 (31,938)	13,274
Avg. # of chars	23.79 (23.80)	23.23

*Numbers in brackets denote the sizes in public part.

4. Experiments — Comparing the performance of image captioning in Japanese

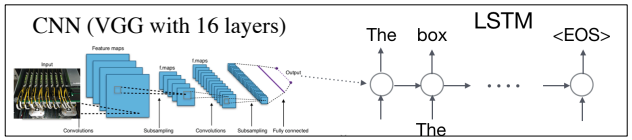
Configuration. We compare two methods using the same neural network (NN) architecture.

- **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-16
- 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.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L	CIDEr
En-generator → MT	0.565	0.330	0.204	0.127	0.449	0.324
Ja-generator	**0.763**	**0.614**	**0.492**	**0.385**	**0.553**	**0.833**

- 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.

En-generator:
A double decker bus driving down a street.
En-generator → MT:
ストリートを運転する二重デッカーバス。 (unnatural)
Ja-generator:
二階建てのバスが道路を走っている。 (natural)

En-generator:
A bunch of food that are on a table.
En-generator → MT:
テーブルの上にある食べ物の束。 (incorrect)
Ja-generator:
ドーナツがたくさん並んでいる。 (correct)