Cross-Lingual Image Caption Generation

Takashi Miyazaki*
Yahoo Japan Corporation
Tokyo, Japan
takmiyaz@yahoo-corp.jp

Nobuyuki Shimizu*
Yahoo Japan Corporation
Tokyo, Japan
nobushim@yahoo-corp.jp

Abstract

Automatically generating a natural language description of an image is a fundamental problem in artificial intelligence. This task involves both computer vision and natural language processing and is called "image caption generation." Research on image caption generation has typically focused on taking in an image and generating a caption in English as existing image caption corpora are mostly in English. The lack of corpora in languages other than English is an issue, especially for morphologically rich languages such as Japanese. There is thus a need for corpora sufficiently large for image captioning in other languages. We have developed a Japanese version of the MS COCO caption dataset and a generative model based on a deep recurrent architecture that takes in an image and uses this Japanese version of the dataset to generate a caption in Japanese. As the Japanese portion of the corpus is small, our model was designed to transfer the knowledge representation obtained from the English portion into the Japanese portion. Experiments showed that the resulting bilingual comparable corpus has better performance than a monolingual corpus, indicating that image understanding using a resource-rich language benefits a resource-poor language.

1 Introduction

Automatically generating image captions by describing the content of an image using natural language sentences is a challenging task. It is especially challenging for languages other than English due to the sparsity of annotated resources in the target language. A promising solution to this problem is to create a comparable corpus. To support the image caption generation task in Japanese, we have annotated images taken from the MS COCO caption dataset (Chen et al., 2015b) with Japanese captions. We call our corpus the "YJ Captions 26k Dataset." While the size of our dataset is comparatively large with 131,740 captions, it greatly trails the 1,026,459 captions in the MS COCO dataset. We were thus motivated to transfer the resources in English (source language) to Japanese and thereby improve image caption generation in Japanese (target language). In natural language processing, a task involving transferring information across languages is known as a cross-lingual natural language task, and well known tasks include cross-lingual sentiment analysis (Chen et al., 2015a), cross-lingual named entity recognition (Zirikly and Hagiwara, 2015), cross-lingual dependency parsing (Guo et al., 2015), and cross-lingual information retrieval (Funaki and Nakayama, 2015).

Existing work in the cross-lingual setting is usually formulated as follows. First, to overcome the language barrier, create a connection between the source and target languages, generally by using a dictionary or parallel corpus. Second, develop an appropriate knowledge transfer approach to leverage the annotated data from the source language for use in training a model in the target language, usually supervised or semi-supervised. These two steps typically amount to automatically generating and expanding the pseudo-training data for the target language by exploiting the knowledge obtained from the source language.

We propose a very simple approach to cross-lingual image caption generation: exploit the English corpus to improve the performance of image caption generation in another language. In this ap-

* Both authors contributed equally to this work.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1780–1790,
Berlin, Germany, August 7-12, 2016.

proach, no resources besides the images found in the corpus are used to connect the languages, and we consider our dataset to be a comparable corpus. Paired texts in a comparable corpus describe the same topic, in this case an image, but unlike a parallel corpus, the texts are not exact translations of each other. This unrestrictive setting enables the model to be used to create image caption resources in other languages. Moreover, this model scales better than creating a parallel corpus with exact translations of the descriptions.

Our transfer model is very simple. We start with a neural image caption model (Vinyals et al., 2015) and pretrain it using the English portion of the corpus. We then remove all of the trained neural network layers except for one crucial layer, the one closest to the vision system. Next we attach an untrained Japanese generation model and train it using the Japanese portion of the corpus. This results in improved generation in Japanese compared to using only the Japanese portion of the corpus. To the best of our knowledge, this is the first paper to address the problem of cross-lingual image caption generation.

Our contribution is twofold. First, we have created and plan to release the first ever significantly large corpus for image caption generation for the Japanese language, forming a comparable corpus with existing English datasets. Second, we have created a very simple model based on neural image caption generation for Japanese that can exploit the English portion of the dataset. Again, we are the first to report results in cross-lingual image caption generation, and our surprisingly simple method improves the evaluation metrics significantly. This method is well suited as a baseline for future work on cross-lingual image caption generation.

The paper is organized as follows. In the next section, we describe related work in image caption generation and list the corpora currently available for caption generation. Then in Section 3 we present the statistics for our corpus and explain how we obtained them. We then explain our model in Section 4 and present the results of our experimental evaluation in Section 5. We discuss the results in Section 6, and conclude in Section 7 with a summary of the key points.

2 Related Work

Recent advances in computer vision research have led to halving the error rate between 2012 and 2014 at the Large Scale Visual Recognition Challenge (Russakovsky et al., 2015), largely driven by the adoption of deep neural networks (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Donahue et al., 2014; Sharif Razavian et al., 2014). Similarly, we have seen increased adaptation of deep neural networks for natural language processing. In particular, sequence-to-sequence training using recurrent neural networks has been successfully applied to machine translation (Cho et al., 2014; Bahdanau et al., 2015; Sutskever et al., 2014; Kalchbrenner and Blunsom, 2013).

These developments over the past few years have led to renewed interest in connecting vision and language. The encoder-decoder framework (Cho et al., 2014) inspired the development of many methods for generating image captions since generating an image caption is analogous to translating an image into a sentence.

Since 2014, many research groups have reported a significant improvement in image caption generation due to using a method that combines a convolutional neural network with a recurrent neural network. Vinyals et al. used a convolutional neural network (CNN) with inception modules for visual recognition and long short-term memory (LSTM) for language modeling (Vinyals et al., 2015). Xu et al. introduced an attention mechanism that aligns visual information and sentence generation for improving captions and understanding of model behavior (Xu et al., 2015). The interested reader can obtain further information elsewhere (Bernardi et al., 2016).

These developments were made possible due to a number of available corpora. The following is a list of available corpora that align images with crowd-sourced captions. A comprehensive list of other kinds of corpora connecting vision and language, e.g., visual question answering, is available elsewhere (Ferraro et al., 2015).

1. UIUC Pascal Dataset (Farhadi et al., 2010) includes 1,000 images with 5 sentences per image; probably one of the first datasets.
2. Abstract Scenes Dataset (Clipart) (Zitnick et al., 2013) contains 10,020 images of children playing outdoors associated with 60,396 descriptions.

3. Flickr 30K Images (Young et al., 2014) extends Flickr datasets (Rashtchian et al., 2010) and contains 31,783 images of people involved in everyday activities.
4. Microsoft COCO Dataset (MS COCO) (Lin et al., 2014; Chen et al., 2015b) includes about 328,000 images of complex everyday scenes with common objects in naturally occurring contexts. Each image is paired with five captions.
5. Japanese UIUC Pascal Dataset (Funaki and Nakayama, 2015) is a Japanese translation of the UIUC Pascal Dataset.

To the best of our knowledge, there are no large datasets for image caption generation except for English. With the release of the YJ Captions 26k dataset, we aim to remedy this situation and thereby expand the research horizon by exploiting the availability of bilingual image caption corpora.

3 Statistics for Data Set

In this section we describe the data statistics and how we gathered data for the YJ Captions 26k dataset. For images, we used the Microsoft COCO dataset (Chen et al., 2015b). The images in this dataset were gathered by searching for pairs of 80 object categories and various scene types on Flickr. They thus tended to contain multiple objects in their natural context. Objects in the scene were labeled using per-instance segmentations. This dataset contains pictures of 91 basic object types with 2.5 million labeled instances. To collect Japanese descriptions of the images, we used Yahoo! Crowdsourcing [1], a microtask crowdsourcing service operated by Yahoo Japan Corporation.

Given 26,500 images taken from the training part of the MS COCO dataset, we collected 131,740 captions in total. The images had on average 4.97 captions; the maximum number was 5 and the minimum was 3. On average, each caption had 23.23 Japanese characters. We plan to release the YJ Captions 26k dataset [2].

3.1 Crowdsourcing Procedure

Our captions were human generated using Yahoo! Crowdsourcing. As this crowdsourcing platform is operated in Japan, signing up for the service and participating require Japanese proficiency. Thus, we assumed that the participants were fluent in Japanese.

Figure 1: User Interface

First, we posted a pilot task that asked the participants to describe an image. We then examined the results and selected promising participants (comprising a "white list") for future task requests. That is, only the participants on the white list could see the next task. This selection process was repeated, and the final white list included about 600 participants. About 150 of them regularly participated in the actual image caption collection task. We modified the task request page and user interface on the basis of our experience with the pilot task. In order to prevent their fatigue, the tasks were given in small batches so that the participants were unable to work over long hours.

In our initial trials, we tried a direct translation of the instructions used in the MS-COCO English captions. This however did not produce Japanese captions comparable to those in English. This is because people describe what appears unfamiliar to them and do not describe things they take for granted. Our examination of the results from the pilot tasks revealed that the participants generally thought that the pictures contained non-Japanese people and foreign places since the images originated from Flickr and no scenery from Japan was included in the image dataset. When Japanese

[1] http://crowdsourcing.yahoo.co.jp
[2] http://research-lab.yahoo.co.jp/software/index.html

crowds are shown pictures with scenery in the US or Europe in MS-COCO dataset, the scenes themselves appear exotic and words such as 'foreign' and 'oversea' would be everywhere in the descriptions. As such words are not common in the original dataset, and to make the corpus nicer complement to the English dataset and to reduce the effects of such cultural bias, we modified the instructions: "2. Please give only factual statements"; "3. Please do not specify place names or nationalities." We also strengthened two sections in the task request page and added more examples.

The interface is shown in Figure 1. The instructions in the user interface can be translated into English as "Please explain the image using 16 or more Japanese characters. Write a single sentence as if you were writing an example sentence to be included in a textbook for learning Japanese. Describe all the important parts of the scene; do not describe unimportant details. Use correct punctuation. Write a single sentence, not multiple sentences or a phrase."

Potential participants are shown task request pages, and the participants select which crowdsourcing task(s) to perform. The task request page for our task had the following instructions (English translation):

1. Please explain an image using 16 or more Japanese characters. Please write a single sentence as if you were writing an example sentence to be included in a textbook for learning Japanese.
   (a) Do not use incorrect Japanese.
   (b) Use a polite style of speech (*desu*/*masu* style) as well as correct punctuation.
   (c) Write a single complete sentence that ends with a period. Do not write just a phrase or multiple sentences.
2. Please give only factual statements.
   (a) Do not write about things that might have happened or might happen in the future. Do not write about sounds.
   (b) Do not speculate. Do not write about something about which you feel uncertain.
   (c) Do not state your feelings about the scene in the picture. Do not use an overly poetic style.
   (d) Do not use a demonstrative pronoun such as 'this' or 'here.'
3. Please do not specify place names or nationalities.
   (a) Please do not give proper names.
4. Please describe all the important parts of the scene; do not describe unimportant details.

Together with the instructions, we provided 15 examples (1 good example; 14 bad examples).

Upon examining the collected data, manual checks of first 100 images containing 500 captions revealed that 9 captions were clearly bad, and 12 captions had minor problems in descriptions. In order to further improve the quality of the corpus, we crowdsourced a new data-cleaning task. We showed each participant an image and five captions that describe the image and asked to fix them.

The following is the instructions (English translation) for the task request page for our data-cleaning task.

1. There are five sentences about a hyper-linked image, and several sentences require fixes in order to satisfy the conditions below. Please fix the sentences, and while doing so, tick a checkbox of the item (condition) being fixed.
2. The conditions that require fixes are:
   (a) Please fix typographical errors, omissions and input-method-editor conversion misses.
   (b) Please remove or rephrase expressions such as 'oversea', 'foreign' and 'foreigner.'
   (c) Please remove or rephrase expressions such as 'image', 'picture' and 'photographed.'
   (d) Please fix the description if it does not match the contents of the image.
   (e) Please remove or rephrase subjective expressions and personal impressions.
   (f) If the statement is divided into several sentences, please make it one sentence.
   (g) If the sentence is in a question form, please make it a declarative sentence.
   (h) Please rewrite the entire sentence if meeting all above conditions requires extensive modifications.
   (i) If there are less than 16 characters, please provide additional descriptions so that the sentence will be longer than 16 characters.

For each condition, we provided a pair of examples (1 bad example and 1 fixed example).

To gather participants for the data-cleaning task, we crowdsourced a preliminary user qualification task that explained each condition requiring fixes in the first half, then quizzed the participants in the second half. This time we obtained over 900 qualified participants. We posted the data-cleaning task to these qualified participants.

The interface is shown in Figure 2. The instructions in the user interface are very similar to the task request page, except that we have an additional checkbox:

(j) All conditions are satisfied and no fixes were necessary.

We provided these checkboxes to be used as a checklist, so as to reduce failure by compensating for potential limits of participants' memory and attention, and to ensure consistency and completeness in carrying out the data-cleaning task.

For this data-cleaning task, we had 26,500 images totaling 132,500 captions checked by 267 participants. The number of fixed captions are

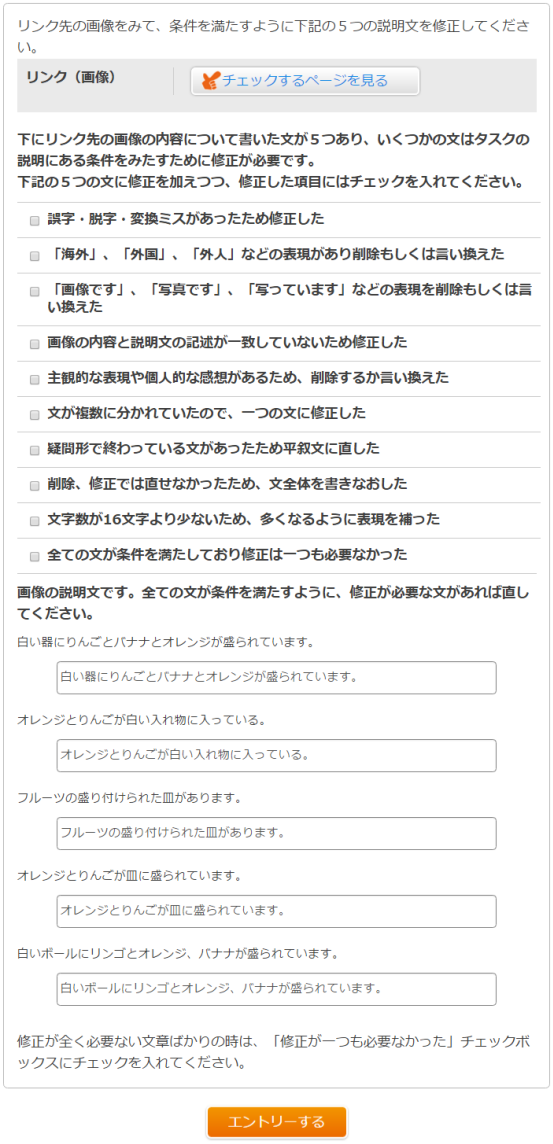

Figure 2: Data Cleaning Task User Interface

45,909. To our surprise, a relatively large portion of the captions were fixed by the participants. We suspect that in our data-cleaning task, the condition (e) was especially ambiguous for the participants, and they errored on the cautious side, fixing "a living room" to just "a room", thinking that a room that looks like a living room may not be a living room for the family who occupies the house, for example. Another example includes fixing "beautiful flowers" to just "flowers" because beauty is in the eye of the beholder and thought to be subjective. The percentage of the ticked checkboxes is as follows: (a) 27.2%, (b) 5.0%, (c) 12.3%, (d) 34.1%, (e) 28.4%, (f) 3.9%, (g) 0.3%, (h) 11.6%, (i) 18.5%, and (j) 24.0%. Note that a checkbox is ticked if there is at least one sentence out of five that meets the condition. In machine learning, this setting is called multiple-instance multiple-label problem (Zhou et al., 2012). We cannot directly infer how many captions correspond to a condition ticked by the participants.

After this data-cleaning task, we further removed a few more bad captions that came to our attention. The resulting corpus finally contains 131,740 captions as noted in the previous section.

4 Methodology

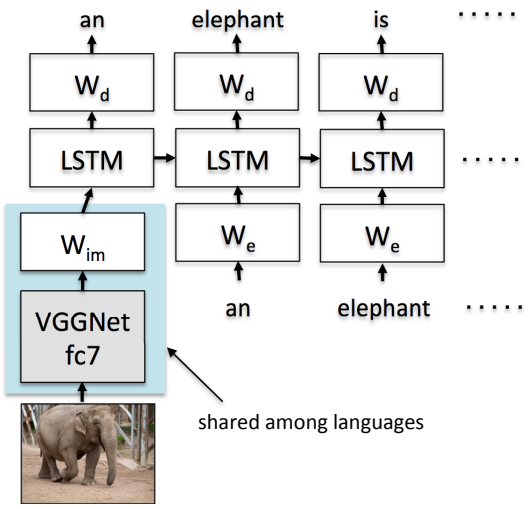

Figure 3: Model Overview

4.1 Model Overview

Figure 3 shows an overview of our model. Following the approach of Vinyals et al. (Vinyals et al., 2015), we used a discriminative model that maximizes the probability of the correct description given the image. Our model is formulated as

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \sum_{t=0}^{N} \log p(S_t|I, S_0, ..., S_{t-1}; \theta), \quad (1)$$

where the first summation is over pairs of an image I and its correct transcription S. For the second summation, the sum is over all words S_t in S, and N is the length of S. θ represents the model parameters. Note that the second summation represents the probability of the sentence with respect to the joint probability of its words.

We modeled $p(S_t|I, S_0, ..., S_{t-1}; \theta)$ by using a recurrent neural network (RNN). To model the sequences in the RNN, we let a fixed length hidden state or memory h_t express the variable number of words to be conditioned up to $t-1$. The h_t

is updated after obtaining a new input x_t using a non-linear function f, so that $h_{t+1} = f(h_t, x_t)$. Since an LSTM network has state-of-the art performance in sequence modeling such as machine translation, we use one for f, which we explain in the next section.

A combination of LSTM and CNN are used to model $p(S_t|I, S_0, ..., S_{t-1}; \theta)$.

$$x_{-1} = W_{im}CNN(I) \quad (2)$$

$$x_t = W_e S_t, t \in \{0...N-1\} \quad (3)$$

$$p_{t+1} = Softmax(W_d LSTM(x_t)), \quad t \in \{0...N-1\} \quad (4)$$

where W_{im} is an image feature encoding matrix, W_e is a word embedding matrix, and W_d is a word decoding matrix.

4.2 LSTM-based Language Model

An LSTM is an RNN that addresses the vanishing and exploding gradients problem and that handles longer dependencies well. An LSTM has a memory cell and various gates to control the input, the output, and the memory behaviors. We use an LSTM with input gate i_t, input modulation gate g_t, output gate o_t, and forgetting gate f_t. The number of hidden units h_t is 256. At each time step t, the LSTM state c_t, h_t is as follows:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (5)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (6)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (7)$$

$$g_t = \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (8)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (9)$$

$$h_t = o_t \odot \phi(c_t), \quad (10)$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is a sigmoid function, $\phi(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is a hyperbolic tangent function, and $\odot$ denotes the element-wise product of two vectors. W and b are parameters to be learned. From the values of the hidden units h_t, the probability distribution of words is calculated as

$$p_{t+1} = Softmax(W_d h_t). \quad (11)$$

We use a simple greedy search to generate captions as a sequence of words, and, at each time step t, the predicted word is obtained using $S_t = \arg\max_S p_t$.

4.3 Image Feature Extraction with Deep Convolutional Neural Network

The image recognition performance of deep convolutional neural network models has rapidly advanced in recent years, and they are now widely used for various image recognition tasks. We used a 16-layer VGGNet (Simonyan and Zisserman, 2014), which was a top performer at the ImageNet Large Scale Visual Recognition Challenge in 2014. A 16-layer VGGNet is composed of 13 convolutional layers having small 3x3 filter kernels and 3 fully connected layers. An image feature is extracted as a 4096-dimensional vector of the VGGNet's fc7 layer, which is the second fully connected layer from the output layer. VGGNet was pretrained using the ILSVRC2014 subset of the ImageNet dataset, and its weights were not updated through training.

4.4 Dataset Split

Because our caption dataset is annotated for only 26,500 images of the MS COCO training set, we reorganized the dataset split for our experiments. Training and validation set images of the MS COCO dataset were mixed and split into four blocks, and these blocks were assigned to training, validation, and testing as shown in Table 1. All blocks were used for the English caption dataset. Blocks B, C, and D were used for the Japanese caption dataset.

block	no. of images	split	language
A	96,787	train	En
B	22,500	train	En, Ja
C	2,000	val	En, Ja
D	2,000	test	En, Ja
total	123,287		

Table 1: Dataset Split

4.5 Training

The models were trained using minibatch stochastic gradient descent, and the gradients were computed by backpropagation through time. Parameter optimization was done using the RMSprop algorithm (Tieleman and Hinton, 2012) with an initial learning rate of 0.001, a decay rate of 0.999, and ϵ of 1.0^{-8}. Each image minibatch contained 100 image features, and the corresponding caption minibatch contained one sampled caption per image. To evaluate the effectiveness of Japanese

image caption generation, we used three learning schemes.

Monolingual learning This was the baseline method. The model had only one LSTM for Japanese caption generation, and only the Japanese caption corpus was used for training.

Alternate learning In this scheme, a model had two LSTMs, one for English and one for Japanese. The training batches for captions contained either English or Japanese, and the batches were fed into the model alternating between English and Japanese.

Transfer learning A model with one LSTM was trained completely for the English dataset. The trained LSTM was then removed, and another LSTM was added for Japanese caption generation. W_{im} was shared between the English and Japanese training.

These models were implemented using the Chainer neural network framework (Tokui et al., 2015). We consulted NeuralTalk (Karpathy, 2014), an open source implemenation of neural network based image caption generation system, for training parameters and dataset preprocessing. Training took about one day using NVIDIA TITAN X/Tesla M40 GPUs.

5 Evaluation

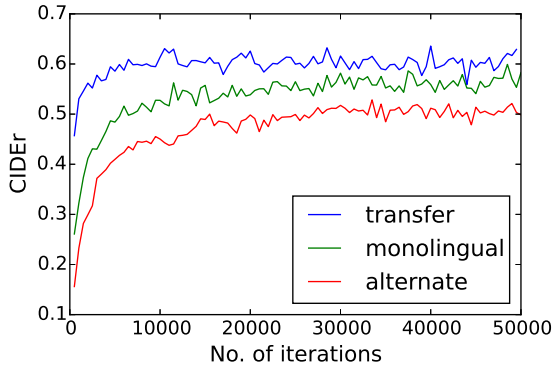

Figure 4: Learning Curve Represented by CIDEr Score

5.1 Evaluation Metrics

We used six standard metrics for evaluating the quality of the generated Japanese sentences: BLEU-1, BLEU-2, BLEU-3, BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr-D (Vedantam et al., 2014). We used the COCO caption evaluation tool (Chen et al., 2015b) to compute the metrics. BLEU (Papineni et al., 2002) was originally designed for automatic machine translation. By counting n-gram co-occurrences, it rates the quality of a translated sentence given several reference sentences. To apply BLEU, we considered that generating image captions is the same as translating images into sentences. ROUGE (Lin, 2004) is an evaluation metric designed by adapting BLEU to evaluate automatic text summarization algorithms. ROUGE is based on the longest common subsequences instead of n-grams. CIDEr (Vedantam et al., 2014) is a metric developed specifically for evaluating image captions. It measures consensus in image captions by performing a term-frequency inverse document frequency (TF-IDF) weighting for each n-gram. We used a robust variant of CIDEr called CIDEr-D. For all evaluation metrics, higher scores are better. In addition to these metrics, MS COCO caption evaluation (Chen et al., 2015b) uses METEOR (Lavie, 2014), another metric for evaluating automatic machine translation. Although METEOR is a good metric, it uses an English thesaurus. It was not used in our study due to the lack of a thesaurus for the Japanese language.

The CIDEr and METEOR metrics perform well in terms of correlation with human judgment (Bernardi et al., 2016). Although BLEU is unable to sufficiently discriminate between judgments, we report the BLEU figures as well since their use in literature is widespread. In the next section, we focus our analysis on CIDEr.

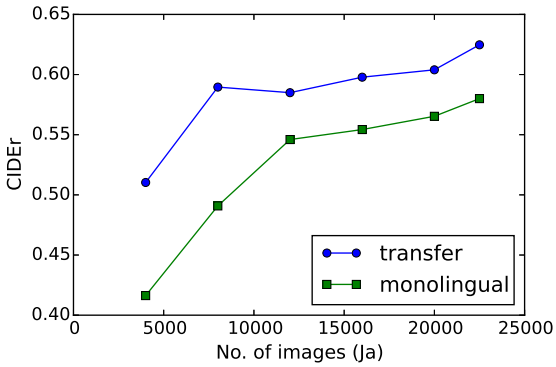

Figure 5: CIDEr Score vs. Japanese Data Set Size

5.2 Results

Table 2 shows the evaluation metrics for various settings of cross-lingual transfer learning. All values were calculated for Japanese captions gener-

	no. of images		metrics					
	En	Ja	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
monolingual	0	22,500	0.715	0.573	0.468	0.379	0.616	0.580
alternate	119,287	22,500	0.709	0.565	0.460	0.370	0.611	0.568
transfer	119,287	22,500	**0.717**	**0.574**	**0.469**	**0.380**	**0.619**	**0.625**

Table 2: Evaluation Metrics

ated for test set images. Our proposed model is labeled "transfer." As you can see, it outperformed the other two models for every metric. In particular, the CIDEr-D score was about 4% higher than that for the monolingual baseline. The performance of a model trained using the English and Japanese corpora alternately is shown on the line label "alternate." Surprisingly, this model had lower performance than the baseline model.

In Figure 4, we plot the learning curves represented by the CIDEr score for the Japanese captions generated for the validation set images. Transfer learning from English to Japanese converged faster than learning from the Japanese dataset or learning by training from both languages alternately. Figure 5 shows the relationship between the CIDEr score and the Japanese dataset size (number of images). The models pretrained using English captions (blue line) outperformed the ones trained using only Japanese captions for all training dataset sizes. As can be seen by comparing the case of 4,000 images with that of 20,000 images, the improvement due to cross-lingual transfer was larger when the Japanese dataset was smaller. These results show that pretraining the model with all available English captions is roughly equivalent to training the model with captions for 10,000 additional images in Japanese. This, in our case, nearly halves the cost of building the corpus.

Examples of machine-generated captions along with the crowd-written ground truth captions (English translations) are shown in Figure 6.

6 Discussion

Despite our initial belief, training by alternating English and Japanese input batch data for learning both languages did not work well for either language. As Japanese is a morphologically rich language and word ordering is subject-object-verb, it is one of most distant languages from English. We suspect that the alternating batch training interfered with learning the syntax of either language. Moreover, when we tried character-based models for both languages, the performance was significantly lower. This was not surprising because one word in English is roughly two characters in Japanese, and presumably differences in the language unit should affect performance. Perhaps not surprisingly, cross-lingual transfer was more effective when the resources in the target language are poor. Convergence was faster with the same amount of data in the target language when pretraining in the source language was done ahead of time. These two findings ease the burden of developing a large corpus in a resource poor language.

7 Conclusion

We have created an image caption dataset for the Japanese language by collecting 131,740 captions for 26,500 images using the Yahoo! Crowdsourcing service in Japan. We showed that pretraining a neural image caption model with the English portion of the corpus improves the performance of a Japanese caption generation model subsequently trained using Japanese data. Pretraining the model using the English captions of 119,287 images was roughly equivalent to training the model using the captions of 10,000 additional images in Japanese. This, in our case, nearly halves the cost of building a corpus. Since this performance gain is obtained without modifying the original monolingual image caption generator, the proposed model can serve as a strong baseline for future research in this area. We hope that our dataset and proposed method kick start studies on cross-lingual image caption generation and that many others follow our lead.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representation (ICLR)*.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank.

Figure 6: Image Caption Generation Examples

2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *arXiv preprint arXiv:1601.03896*.

Qiang Chen, Wenjie Li, Yu Lei, Xule Liu, and Yanxiang He. 2015a. Learning to adapt credible knowledge in cross-lingual sentiment analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 419–429, Beijing, China, July. Association for Computational Linguistics.

Xinlei Chen, Tsung-Yi Lin Hao Fang, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollr, and C. Lawrence Zitnick. 2015b. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference in Machine Learning (ICML)*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 15–29, Berlin, Heidelberg. Springer-Verlag.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Lisbon, Portugal, September. Association for Computational Linguistics.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 585–590, Lisbon, Portugal, September. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China, July. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October. Association for Computational Linguistics.

Andrej Karpathy. 2014. Neuraltalk. `https://github.com/karpathy/neuraltalk`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Michael Denkowski Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. *ACL 2014*, page 376.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, 2014. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, chapter Microsoft COCO: Common Objects in Context, pages 740–755. Springer International Publishing, Cham.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text summarization branches out: Proceedings of the ACL-04 workshop*, 8.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. Cnn features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June.

K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

T. Tieleman and G. Hinton. 2012. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. 2012. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320.

Ayah Zirikly and Masato Hagiwara. 2015. Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 390–396, Beijing, China, July. Association for Computational Linguistics.

C.L. Zitnick, D. Parikh, and L. Vanderwende. 2013. Learning the visual interpretation of sentences. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1681–1688, Dec.