Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4204–4210
Marseille, 11–16 May 2020

A Visually-Grounded Parallel Corpus with Phrase-to-Region Linking

Hideki Nakayama[1], **Akihiro Tamura**[2], **Takashi Ninomiya**[2]
[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]Graduate School of Science and Engineering, Ehime University
nakayama@ci.i.u-tokyo.ac.jp, {tamura, ninomiya}@cs.ehime-u.ac.jp

Abstract

Visually-grounded natural language processing has become an important research direction in the past few years. However, majorities of the available cross-modal resources (e.g., image-caption datasets) are built in English and cannot be directly utilized in multilingual or non-English scenarios. In this study, we present a novel multilingual multimodal corpus by extending the Flickr30k Entities image-caption dataset with Japanese translations, which we name Flickr30k Entities JP (F30kEnt-JP). To the best of our knowledge, this is the first multilingual image-caption dataset where the captions in the two languages are parallel and have the shared annotations of many-to-many phrase-to-region linking. We believe that phrase-to-region as well as phrase-to-phrase supervision can play a vital role in fine-grained grounding of language and vision, and will promote many tasks such as multilingual image captioning and multimodal machine translation. To verify our dataset, we performed phrase localization experiments in both languages and investigated the effectiveness of our Japanese annotations as well as multilingual learning realized by our dataset.

Keywords: Multimodal Parallel Corpus, Visual Grounding, Phrase-to-Region Matching, Multimodal Machine Translation

1. Introduction

Grounding natural language to other modalities has recently been gaining much attention as a promising approach to go beyond traditional NLP by incorporating information or knowledge not readily available in textual domain. Particularly, visual grounding has become a popular topic because of the richness and importance of visual information in our perception of the world, and many so-called *language and vision* tasks have been proposed. Representative examples include image and video captioning (Vinyals et al., 2015), visual question answering (Agrawal et al., 2015), visual story telling (Huang et al., 2016), visual dialog (Das et al., 2017), and multimodal machine translation (Elliott et al., 2015; Hitschler et al., 2016).

To promote the research based on visual grounding, multimodal corpus, i.e. a set of pairs of an image and a natural language description depicting the same underlying concept, has been an indispensable resource. However, existing multimodal corpora have limitations in the following aspects. First, most of the existing large-scale corpora are monolingual and built in English (Lin et al., 2014; Krishna et al., 2017; Sharma et al., 2018). While they are useful for *language and vision* problems in general, they cannot be directly used for the study of cross-lingual tasks such as machine translation, and specific problems in each non-English language. Second, although there exist some multilingual multimodal corpora that cover more than one language, the link between images and texts is not very strong in the sense that only global-level (i.e., sentence-to-image) correspondence is available. To realize more fine-grained visual grounding in multilingual scenarios, it is desirable to supervise entity-level linking of visual and textual representations in multiple languages.

Based on these motivations, in this study, we propose a multilingual multimodal parallel corpus with the annotation of phrase-to-region linking, which we name Flickr30k Entities JP (F30kEnt-JP). It is a Japanese extension of the original Flickr30K Entities dataset (Plummer et al., 2017), an English image-caption dataset which has phrase-to-region annotations. More specifically, we translate its English captions into Japanese preserving the original phrase-to-region annotations in the translated Japanese captions (Figure 1). As the basic evaluation of the proposed dataset, we performed phrase-to-region retrieval experiment through which we confirmed the effectiveness of the Japanese captions and multilingual learning. We believe F30kEnt-JP can facilitate *language and vision* researches in both multilingual and Japanese-specific contexts[1].

2. Related Work

In this section, we introduce existing multilingual multimodal datasets and discuss the novel contribution of ours. Many researchers have proposed multilingual multimodal datasets by extending some standard English image-caption datasets such as Pascal Sentences (Rashtchian et al., 2010), Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014), MS-COCO (Lin et al., 2014), and Visual Genome (Krishna et al., 2017). For example, Chinese extensions of Flickr8k (Li et al., 2016) and Flickr30k (Lan et al., 2017), Japanese extensions of MS-COCO (Miyazaki and Shimizu, 2016; Yoshikawa et al., 2017) have been proposed. Similarly, some subsets of the MS-COCO dataset are extended with Chinese (Li et al., 2019) and German (Hitschler et al., 2016). The primary target in these datasets is multilingual image captioning. Because the captions in another language are annotated to the images independently from the original English captions, they are thought to be somewhat comparable but not strictly parallel [2].

[1]The F30kEnt-JP dataset is publicly available at: `https://github.com/nlab-mpg/Flickr30kEnt-JP`.

[2]Although the testing sets of Flickr8k-CN, Flickr30k-CN and COCO-CN contain manually translated Chinese captions, they are

Dataset	Languages	Source	# Imgs	# Sent.	Human Trans.	Phrase-Region Linking
IAPR-TC12 (Henning et al., 2006)	German, English	Web images	20,000	20,000	Yes	No
Pascal Sentences JP (Funaki and Nakayama, 2015)	Japanese, English	Pascal Sentences	1,000	5,000	Yes	No
Flickr8k-CN (Li et al., 2016)	Chinese, English	Flickr8k	8,000	40,000	Partial (test set)	No
Flickr30k-CN (Lan et al., 2017)	Chinese, English	Flickr30k	31,783	158,915	Partial (test set)	No
Multi30K (Translations) (Elliott et al., 2016)	German, French, Czech, English	Flickr30k	31,014	31,014	Yes	No
YJ Captions (Miyazaki and Shimizu, 2016)	Japanese, English	MS-COCO	26,500	131,740	No	No
STAIR Captions (Yoshikawa et al., 2017)	Japanese, English	MS-COCO	164,062	820,310	No	No
MIC test data (Rajendran et al., 2016)	German, French, English	MS-COCO	1,000	5,000	No	No
Bilingual caption (Hitschler et al., 2016)	German, English	MS-COCO	1,000	1,000	Yes	No
COCO-CN (Li et al., 2019)	Chinese, English	MS-COCO	20,342	27,218	Partial (test set)	No
Hindi Visual Genome (Parida et al., 2019)	Hindi, English	Visual Genome	31,525	31,525	Yes	Yes (one-to-one)
Flickr30k Entities JP (Ours)	**Japanese, English**	**Flickr30k Entities**	**31,783**	**63,566**	**Yes**	**Yes**

Table 1: Summary of existing multilingual image-caption datasets and ours.

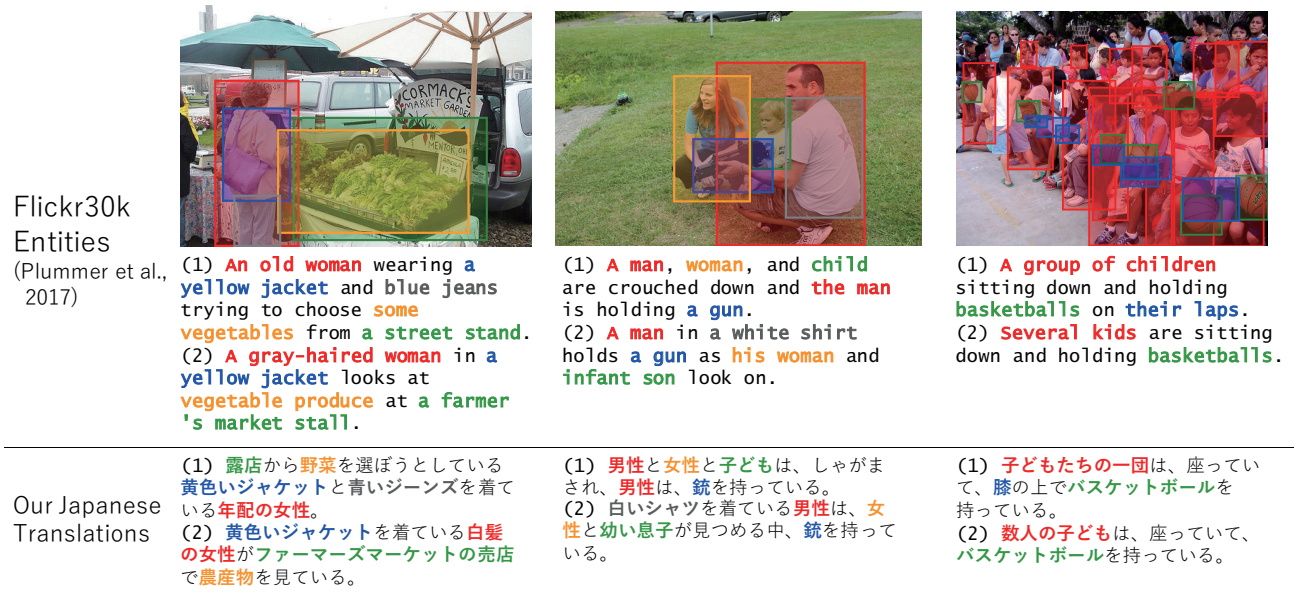


Figure 1: Examples from Flickr30k Entities and our Japanese translations (Flickr30k Entities JP). Corresponding phrases and image regions are highlighted with the same color. (This figure is best viewed in color.)

Meanwhile, some researchers have provided multimodal corpora that have strictly parallel captions. IAPR-TC12 (Henning et al., 2006) is probably the first dataset in this category. It consists of 20,000 images collected from Web which were originally annotated with German captions. It also has English captions manually translated from the German ones. Pascal Sentence Japanese (Funaki and Nakayama, 2015) is an extension of the Pascal Sentence dataset with Japanese captions obtained by manually translating the original English captions. Multi30K (Elliott et al., 2016) is currently the largest dataset in this context and has been the benchmark resource for multi-

relatively small (1k images each) and mainly used for evaluation purposes.

modal machine translation tasks in WMT workshops [3]. It is mainly based on Flickr30k and has 31,014 images, each of which has one set of parallel captions. As of WMT 2018, it has German, French and Czech captions translated from the original English captions (Barrault et al., 2018). Hindi Visual Genome (Parida et al., 2019) is probably the closest dataset to ours in concepts. It extends the Visual Genome dataset (Krishna et al., 2017) with Hindi translations obtained by NMT and human post-editing. In consists of 31,525 images, and each image has a pair of English and Hindi captions that are linked to a relevant region (bounding box) in the image. As the entire caption is linked to one image region, it makes one-to-one mapping between a sentence (or a phrase) and an image. One notable difference in our dataset is that multiple phrases in a sentence are respectively linked to corresponding image regions (Figure 1). We believe this many-to-many mapping within a sentence and an image is useful to help more structural understanding of natural languages and images.

We summarize the details of the existing datasets and our F30kEnt-JP in Table 1. Notable properties of our dataset are as follows.

- The largest Japanese and English image-caption dataset where the captions in the two languages are parallel.
- The first multilingual multimodal dataset with the annotation of many-to-many region-to-phrase linking as well as phrase-to-phrase linking.

3. Details of the Proposed Dataset

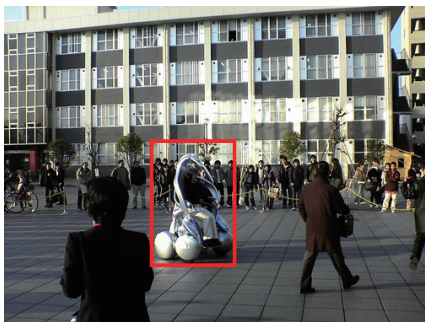
3.1. Source Dataset

The origin of our dataset is the Flickr30k (Young et al., 2014) which was proposed for the study of image captioning. It contains 31,783 images where each image is annotated with five sentences (captions). To facilitate more local-level visual grounding, Plummer et al. (2017) extended Flickr30k by additionally annotating phrase-to-region linking for some phrases in each sentence as shown in Figure 1. We further extend the Flickr30k Entities by carefully translating the original English captions into Japanese keeping the phrase-to-region correspondence. Until now, we have translated the first two captions for each image, resulting in 63,566 parallel sentences in total.

3.2. Construction of the Dataset

English to Japanese translation was done by a professional translation company in Japan which has rich experience in constructing English-Japanese parallel corpora for machine translation research. They were asked to follow two rules while they translated the English sentences. First, the image corresponding to a source sentence (caption) must be referred. This is important because the captions are sometimes less informative and ambiguous, and their meaning cannot be correctly conveyed without seeing the images. Consider the image and English caption in Figure 2, for example. If we only see the source caption, it would be difficult to clearly identify whether the phrase “a weird vehicle” is talking about an automobile (自動車) or a general vehicle (乗り物). By also seeing the corresponding image, we can easily resolve this kind of ambiguity, which indeed is the motivation of multimodal machine translation. Second, English phrases originally linked to image regions must also appear in the Japanese sentence as far as possible. In other words, we expect to have the triplets of the English phrase, Japanese phrase and image region for the entities annotated in the original Flickr30k Entities [4].

However, because of the linguistic discrepancy between Japanese and English, sometimes it is impossible or not natural to put a Japanese phrase that explicitly corresponds to the original English one. We list examples of some representative cases in Figure 3. (a) Pronouns and relative pronouns are often abbreviated in Japanese. (b) Idiomatic discrepancy: the English idiom “side by side” is commonly translated to “並んで” in Japanese, making it meaningless to consider the correspondence to each “side”. (c) Split correspondence: it is more natural to translate “Olympics in Beijing” into “北京オリンピック” (北京 =Beijing, オリンピック =Olympics), resulting in the separation of “the 2008” (2008 年) and “Olympics” (オリンピック) in the translated Japanese caption. In these difficult cases, we simply ignore the original English phrase. Therefore, some phrases in original English captions are inevitably lost in our Japanese translations. We summarize the statistics of the dataset in Table 2. We can see that the loss mostly occurs in non-visual phrases.



En (Original) People are watching a person in a weird vehicle in a plaza.

Jp 人々が、広場で奇妙な乗り物に乗っている人をじっと見ている。

Figure 2: An example where the image can help resolve ambiguity and correct translation of a phrase.

3.3. Possible Applications

We believe that our dataset can be utilized in many scenarios. The primary target would be multilingual image

[3] Multi30K also has annotations for multilingual image captioning task where each image is enriched with five non-parallel German captions.

[4] Some phrases in the original Flickr30k Entities are not linked to image regions, which are marked as *nonvisual*. For those phrases, we just try to keep phrase-to-phrase correspondence in translation (see Table 2).

(a) 

[A casually dressed man]¹ has [several flutes]² into [which]³ [he]⁴ is pouring [champagne]⁵.

[カジュアルな服装の男性]¹は[いくつかのシャンパングラス]²を持っており、そこへ[シャンパン]⁵を注いでいる。

(b) 

[Two guys]¹ standing [side]² by [side]³.

並んで立っている[二人]¹。

(c) 

[Visitors]¹ pose for [a picture]² at [the 2008 Olympics]³ in [Beijing]⁴.

２００８年の[北京]⁴オリンピックで、[観光客]¹が[写真]²のためにポーズを取っている。

Figure 3: Examples where it is difficult to explicitly align some phrases. Ignored phrases in original English captions are underlined.

	Language	# Phrases	# Visually-grounded Phrases
Train	En	263,562	244,594
	Jp	253,722	243,431
Val	En	8,903	8,247
	Jp	8,565	8,214
Test	En	8,792	8,158
	Jp	8,492	8,131

Table 2: Numbers of annotated phrases appearing in our dataset.

captioning and multimodal machine translation, where our phrase-to-region annotations will enable focusing on locally important regions to extract more detailed information from images. Another interesting task is region-level visual understanding such as phrase localization and visual reference expression. Also, unsupervised phrase grounding in multiple languages would be an interesting direction.

In addition to such general multimodal or multilingual problems, our dataset is also useful for Japanese-specific problems. For example, in many Asian languages including Japanese, word segmentation is a critically important step as there is no obvious separation of words (e.g., whitespace) in sentences for those languages. Phrase-level visual grounding could be a useful hint to improve automatic word segmentation.

4. Phrase Localization Experiment

While the overall sentence-level translation quality is maintained by a professional translation team, the quality and effectiveness of phrase-to-region annotation in Japanese should be investigated. To verify this point, we perform the phrase localization experiment, which is proposed as a benchmarking task in the original Flickr30k Entities. Namely, given a phrase in a ground truth caption as the query, the system should correctly detect the corresponding region in the image.

4.1. Methods

We use canonical correlation analysis (CCA) (Hotelling, 1936) to ground visual and textual information. It has been successfully used in many cross-modal tasks such as image annotation (Hardoon et al., 2004; Nakayama et al., 2010), semantic text-based image retrieval (Gong et al., 2014), and cross-lingual document retrieval (Udupa and Khapra, 2010; Funaki and Nakayama, 2015). While many multimodal alignment methods have been proposed based on deep learning recently, we believe CCA is a reasonable baseline to focus on the evaluation of the dataset itself as it is a simple linear method with a less number of hyperparameters.

Let $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ denote the feature vectors of a Japanese phrase, an image region, and an English phrase, respectively. To realize Japanese phrase localization system, for example, we are given access to the pairs $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ which are available in the training image and caption data. The goal of CCA is to find pairs of linear transformations $s = \boldsymbol{a}^T\boldsymbol{x}$ and $t = \boldsymbol{b}^T\boldsymbol{y}$ so that the correlation of the canonical variables s and t is maximized. They can be analytically obtained as the eigenvectors of the following generalized eigenvalue problem.

$$\begin{pmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix}, \quad (1)$$

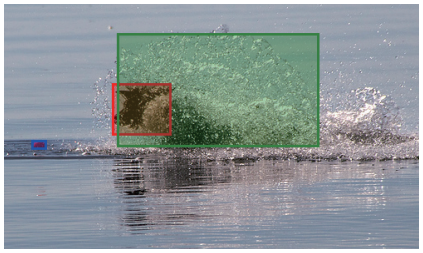
where Σ_{xx} and Σ_{yy} are the variance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively, and $\Sigma_{xy} = \Sigma_{yx}^T$ is their covariance matrix. λ represents the eigenvalue which corresponds to the canonical correlation. We can arbitrarily set the number of the canonical variables by using the eigenvectors in the descending order of their eigenvalues. Once the linear projections are obtained, we can retrieve the most relevant image feature for a query phrase by simply taking the nearest one in the canonical subspace.

Moreover, CCA can be generalized to align more than two modalities, which is called generalized canonical correlation analysis (GCCA). Among several alternatives, Gong et al. (2014) derived the following formulation.

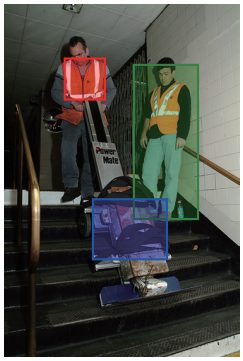
$$\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{xz} \\ \Sigma_{yx} & \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zx} & \Sigma_{zy} & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{pmatrix} = \rho \begin{pmatrix} \Sigma_{xx} & 0 & 0 \\ 0 & \Sigma_{yy} & 0 \\ 0 & 0 & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{pmatrix}, \quad (2)$$

By simultaneously aligning the relevant third modality $\boldsymbol{z}$, we can expect better generalization in a similar flavor of multitask learning. For example, for text-to-image alignment and retrieval problem, Gong et al. (2014) utilized higher-level semantic concepts as the third modality in GCCA and improved the performance from the standard two-view alignment via CCA. In our experiment, we use the English sentence as the third modality and investigate whether this multilingual learning approach can improve phrase localization performance in Japanese (and vice versa).

[シャツを着ていない子ども]が座っている[青いドア]の横で、[小さな少女]が[コンクリートの壁]に寄りかかっている。
(a shirtless child) (a blue door) (a little girl) (a concrete wall)



[犬]が水の中で[大きな水しぶき]を立てながら、[赤いボール]を追いかけている。
(a dog) (a large splash) (a red ball)



[オレンジのベスト]を着ている[二人の男性]は、階段を清掃するために、[工業用清掃機]を使っている。
(orange vest) (two men) (an industrial cleaning machine)

Figure 4: Example results of phrase localization in Japanese. Given a ground truth Japanese caption for an image, we show the region (object proposal) of the top one retrieval for each phrase in the sentence (color corresponds).

4.2. Details of the Experimental Setup

General: We basically follow the same procedure as in (Plummer et al., 2017). We use the splits of 29,783 training, 1,000 validation, and 1,000 testing images provided in the original Flickr30k. Given a test image and a query phrase, we retrieve the closest region from the candidates, which is regarded as the correct match if its intersection of union (IoU) with the ground truth bounding box is more than 0.5. As the evaluation metric, we report Recall@K, i.e. the percentage of the queries that correctly match within top K candidate regions. We ignore non-visual phrases in both training and testing phases. In all experiments, we tune the hyperparameters on the validation dataset.

Region Extraction: For training, we use the ground truth bounding boxes of regions. For validation and testing, we first produce 100 candidate regions using the region proposal network (RPN) bundled in Faster-RCNN object detection (Ren et al., 2015). We use PyTorch vision tools and use the RPN pre-trained on MS-COCO which has ResNet50 as its backbone CNN.

Visual Features: To extract region features, we use VGG16 (Simonyan and Zisserman, 2015) and ResNet50 (He et al., 2016) CNN models pre-trained on ImageNet (Deng et al., 2009) . Specifically, we simply feed forward a cropped region image to the network and take the activation of the last layer before the final fully-connected layer.

Phrase Features: We use the simple average of pre-trained word2vec (Mikolov et al., 2013) embeddings of words in a phrase. For English, we use the 300-dim word vectors pre-trained on the Google News dataset [5]. For Japanese, we first apply the MeCab Japanese tokenizer (Kudo et al., 2004)[6] for word segmentation, and then use the 300-dim word vectors pre-trained on Japanese pages in Wikipedia [7].

It is known that there is a significant imbalance of occurrences of the phrases in the dataset. For example, while common phrases like "a man" frequently appear, minor ones may appear only once in the entire training set. Because taking all examples leads to significantly biased learning and performance drop, we randomly sample at most 10 examples for one phrase as done in (Plummer et al., 2017).

4.3. Results

Table 3 summarizes the evaluation results. Overall, for each method and base CNN, the performance of Japanese phrase localization is slightly lower than that of English by a few percents. Considering the uncertainty in word segmentation and the small vocabulary size of the word vector, we conclude this result is satisfactory to show that phrase-region linking in Japanese is as informative as the original one in English.

Moreover, we observe that GCCA outperforms CCA in most cases. This result indicate that multilingual learning can improve generalization as expected and thus multilingual resource is worth developing. We show some phrase localization results obtained by ResNet50 and GCCA model in Figure 4.

5. Conclusion

In this work, we have developed a novel multilingual multimodal corpus that consists of triplets of Japanese sentences, English sentences and images. We extended the Flickr30k Entities dataset with Japanese sentences by translating the original English captions while keeping the many-to-many phrase-level correspondence both in image regions and the Japanese translations. This annotation makes it possible to

[5] https://code.google.com/archive/p/word2vec/
[6] https://taku910.github.io/mecab/
[7] https://github.com/Kyubyong/wordvectors/

	CNN	Method	R@1	R@5	R@10
En	VGG16	CCA	29.5	45.0	51.6
		GCCA	29.6	44.6	50.6
	ResNet50	CCA	32.0	46.6	53.3
		GCCA	32.3	47.3	53.8
Jp	VGG16	CCA	26.7	40.7	47.3
		GCCA	28.1	41.8	48.4
	ResNet50	CCA	29.8	43.9	50.4
		GCCA	30.2	44.4	51.7

Table 3: Phrase localization performance of the methods trained on the entire corpus. Recall@K (%).

strongly supervise phrase-level matching of texts and images, which is expected to enhance more fine-grained visual grounding and multimodal language processing. We performed the phrase localization experiment as the benchmarking task to investigate the quality and effectiveness of the new Japanese annotations. They are shown to realize comparable localization performance as original English dataset, and further improve the performance when used together with English phrases in the form of multilingual learning.

In future, we would like to use our dataset for more challenging multilingual tasks such as multilingual image captioning and multimodal machine translation. Also, we plan to continue extending the dataset by translating the remaining paraphrases in the original English captions of Flickr30k Entities.

6. Acknowledgements

The results have been achieved by "Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation", the Commissioned Research of National Institute of Information and Communications Technology (NICT), JAPAN.

7. Bibliographical References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., and Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Lee, S., Moura, J. M., Parikh, D., and Batra, D. (2017). Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Elliott, D., Frank, S., and Hasler, E. (2015). Multilanguage Image Description with Neural Sequence Models. In *arXiv preprint arXiv:1510.04709*, pages 1–14.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the ACL 5th Workshop on Vision and Language*, pages 70–74.

Funaki, R. and Nakayama, H. (2015). Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 585–590.

Gong, Y., Ke, Q., Isard, M., and Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233.

Hardoon, D. R., Szedmak, S., and Shawe-taylor, J. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Henning, M., Thomas, D., and Others. (2006). The IAPR Benchmark: ANew Evaluation Resource for Visual Information Systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Hitschler, J., Schamoni, S., and Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2399–2409.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Hotelling, H. (1936). Relations Between Two Sets of Variants. *Biometrika*, 28:321–377.

Huang, T. H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. (2016). Visual storytelling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1233–1239.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L. J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 230–237.

Lan, W., Li, X., and Dong, J. (2017). Fluency-guided cross-lingual image captioning. In *Proceedings of the ACM Multimedia Conference*, pages 1549–1557.

Li, X., Lan, W., Dong, J., and Liu, H. (2016). Adding

Chinese captions to images. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 271–275.

Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., and Xu, J. (2019). COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 1–9.

Miyazaki, T. and Shimizu, N. (2016). Cross-lingual image caption generation. In *Proceedings of th Annual Meeting of the Association for Computational Liguistics (ACL)*, volume 3, pages 1780–1790.

Nakayama, H., Harada, T., and Kuniyoshi, Y. (2010). Evaluation of dimensionality reduction methods for image auto-annotation. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–12.

Parida, S., Bojar, O., and Dash, S. R. (2019). Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *Computación y Sistemas*, 23(4):1499–1505.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*, 123(1):74–93.

Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B. (2016). Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 171–178.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting Image Annotations Using Amazon ' s Mechanical Turk. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2556–2565.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recoginition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Udupa, R. and Khapra, M. (2010). Improving the multilingual user experience of Wikipedia using cross-language name search. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, number June, pages 492–500.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and Tell : A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. (2017). STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 417–421.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations. *Transactions of the Association of Computational Linguistics*, 2(1):67–78.