

Image Description Dataset for Language Learners

Kento Tanaka¹, Taichi Nishimura¹, Hiroaki Nanjo², Keisuke Shirai¹,
Hirotaka Kameko², Masatake Dantsuji³

¹: Graduate School of Informatics, Kyoto University,

²: Academic Center for Computing and Media Studies, Kyoto University

³: Faculty of English and Global Communication, Kyoto Tachibana University

^{1,2}: Yoshidahonmachi, Sakyo-ku, Kyoto, Japan, 606-8501

³: 34 Oyake Yamada, Yamashina-ku, Kyoto, Japan, 607-8175

¹tanaka.kento.s07@kyoto-u.jp, ¹{nishimura.taichi43x,shirai.keisuke.64x}@st.kyoto-u.ac.jp

²nanjo@media.kyoto-u.ac.jp ²kameko@i.kyoto-u.ac.jp ³dantsuji@tachibana-u.ac.jp

Abstract

We focus on image description and a corresponding assessment system for language learners. To achieve automatic assessment of image description, we construct a novel dataset, the Language Learner Image Description (LLID) dataset, which consists of images, their descriptions, and assessment annotations. Then, we propose a novel task of automatic error correction for image description, and we develop a baseline model that encodes multimodal information from a learner sentence with an image and accurately decodes a corrected sentence. Our experimental results show that the developed model can revise errors that cannot be revised without an image.

Keywords: Image Description, Sentence Error Correction, Language Learning

1. Introduction

Language learning involves four basic skills: reading, listening, writing, and speaking. The former two are perception skills, while the latter two are production skills. For developing perception skills, learners are typically given texts or speech and asked to answer multiple-choice questions. Because learners' answers are limited, automatic assessment is not difficult in this situation. In contrast, for developing production skills, learners produce various texts or speech, and an automatic assessment system must handle a diverse range of answers and/or reduce the answer diversity.

Among possible approaches for reducing the answer diversity, one approach is to use a translation task. Given a text written in a learner's native language, the learner is asked to produce the text in the target language. The range of produced texts is expected to be limited by the given texts. However, automatic assessment of the translated texts would require a system that depends on both the learner's native language and the target language. Another possible approach is to use an image description task¹, in which a learner is asked to produce a sentence that describes an image. This task also limits the learner's output to some extent, and it is independent of the learner's native language because the questions consist only of images. Accordingly, this approach would enable learners with different native languages to use the same automatic assessment system for a given target language.

¹Note that this task is usually called "picture description." In the information processing field, however, the term "image" is used instead of "picture," and we thus use the term "image description" instead of "picture description" in this paper.

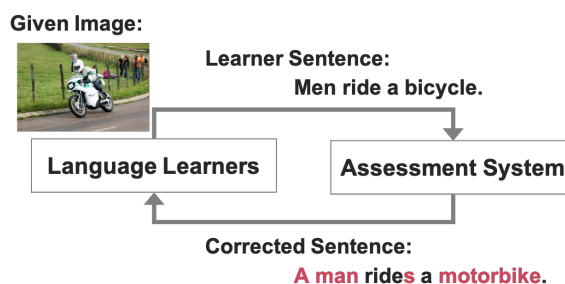


Figure 1: Concept of the assessment system for image description.

Given the above background, in this paper, we focus on image description and a corresponding assessment system, as shown in Figure 1. To achieve automatic assessment of image description texts, we need a dataset that consists of images, their descriptions, and assessment annotations. For the image captioning task, which involves automatic caption generation for a given image, there is much research on neural multimodal models (Vinyals et al., 2015; Xu et al., 2015; Li et al., 2020) using common large-scale datasets such as MS-COCO captions (Lin et al., 2014). These models and common datasets for image captioning can be used to pre-train an assessment system for image description. Accordingly, for this work, we selected specific images from a common image caption dataset; we then collected image description texts by English learners and had them corrected by English experts. The result was a novel dataset for machine learning, which is specifically useful for automatic assessment of image description texts. We call the new dataset the "Language Learner Image Description (LLID) dataset".

In this paper, we first describe how the LLID dataset was constructed (Section 3) and how learners’ errors were analyzed (Section 4). Then, we introduce baseline error correction models for image description texts (Section 5). Finally, we present experimental results and clarify the baseline performance and future problems (Section 6).

The main contributions of this study are as follows:

- We constructed a new dataset that is suitable for automatic assessment of image description.
- We proposed a novel task of automatic error correction for image description, developed a baseline model, and evaluated the model’s performance.

2. Related Work

2.1. Language Learning

Computer-assisted language learning (CALL) has long been investigated as a means to develop learners’ skills. Automatic assessment of learners’ skills is a significant topic in CALL, and many studies have investigated such assessment for the speaking (Duan et al., 2020) and writing (Ng et al., 2014; Bryant et al., 2019) skills. One of the most popular CALL platforms is Duolingo², which is available online. A shared task, the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (Settles et al., 2018), was recently launched. The corresponding dataset includes Duolingo users’ data for the three tasks of “reverse translate”, “reverse tap”, and “listen”. A learner’s error is automatically detected by matching the learner’s sentence to the most similar reference among many possible answers (Settles et al., 2018).

In an image description task, however, many more answers may be acceptable than in the above tasks, because the problem of describing an image’s content is subjective and depends on the learner. This makes it difficult to prepare all possible answers beforehand to enable automatic assessment. On the other hand, questions for image description do not require any prompts in the learner’s native language but only images. Therefore, the image description task is both challenging and promising.

2.2. Grammatical Error Correction (GEC)

Our proposed task of automatic error correction for image description is similar to grammatical error correction (GEC), which aims to correct grammatical errors in a sentence. This is a popular topic in natural language processing (NLP), and it has been investigated extensively (Zhao et al., 2019; Omelanchuk et al., 2020).

Even if perfect GEC were achieved, the image description task would still involve errors that could not be corrected even by a perfect GEC system. Specifically,

GEC cannot correct a sentence that has semantic or pragmatic errors but no grammatical errors. For example, the sentence, “There is a bench on glass,” has no grammatical errors but does have a substitution error, from “grass” to “glass.” In contrast, we assume the use of an image to correct “glass” to “grass.” In addition to such mistakes, GEC cannot correct preposition errors. For example, “in NOUN” and “on NOUN” are both correct English phrases. In this case, the correct preposition cannot be selected without an image.

As described above, with only GEC models, learners do not receive feedback and cannot recognize such errors in their produced sentences, which prevents further development of their language skills. As detailed below, automatic error correction in the image description task can solve these problems.

3. Construction of LLID dataset

We constructed the LLID dataset by (1) asking language learners to answer image description exercises in English and then (2) having them corrected by English experts. The following subsections describe these two processes in detail.

3.1. Data Collection

We conducted the image description exercises with Japanese high school students who were learning English. Given an image and a specific object in the image, the learners were asked to write one sentence in English to describe the object by following our instructions, as shown in Figure 2.

We also collected answers from anonymous learners on the web using our developed web tool as shown in Figure 3.

For an image resource, we used RefCOCOg (Mao et al., 2016), which consists of 85,474 sentences for 54,822 objects in 26,711 images from MS-COCO. Because the images in RefCOCOg are annotated with multiple objects, we manually selected one object per image without biasing the image categories, such as animals, foods, and sports.

As a result, we collected 651 sentences for 120 images. The averaged number of collected descriptions is 5.43 for each image, ranging from 2 to 9. The first line of Table 1 lists the statistics of the collected sentences.³

3.2. Expert Correction

Next, we asked English experts to annotate two types of corrections to the collected sentences: (i) “correctness edits” to correct errors in sentences and (ii) “engagement edits” to give sentences more detailed information about the corresponding image. Examples of these corrections are shown in Figure 4.

³We asked participants to do self-assessment of their proficiency level based on Common European Framework of Reference for Languages (CEFR (Council of Europe, 2001)). As a result, over half of them (60.3%) are A1 level learners.

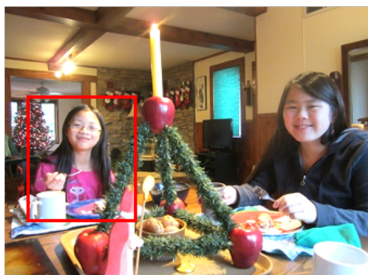
²<https://www.duolingo.com>

Question:

Describe in one sentence in English what the object (person, animal, food, etc.) in the image below is. Note that you must explain a detailed description of where the object is, what the situation is, and what it is doing.

- Sentences should consist of at least **4 words**.
- Do not include anything difficult to determine from the image (e.g., speculation about the future or past).
- You may describe anything outside the frame too, however, be sure to mention the information within the frame.
- Do not include your personal information.
- Do not use your dictionary and reference books.

Example:



Correct Examples

- An Asian girl with a pink shirt is eating at the table next to a woman with black clothes. (Includes information outside the frame)
- A girl is wearing glasses and a pink shirt.

Incorrect Examples

- A girl smiles. (Less than 4 words)
- She will go to school after lunch. (Includes guesses about the future)
- A girl likes her sister. (Includes information difficult to determine from the image)

Figure 2: Instructions for image description. Note that all instructions are originally written in Japanese, but we translate them into English in this paper.

	# Sentences	# Words per Sentence	# Unique Words
Learner Sentences	651	7.12	951
Correctness Edits	644	7.54	654
Engagement Edits	651	14.34	986

Table 1: Statistics of the constructed LLID dataset.

Question3

Describe in one sentence in English what the object (person, animal, food, etc.) in the image below is. Note that you must explain a detailed description of where the object is, what the situation is, and what it is doing.



http://farm9.staticflickr.com/8338/821905538_0918791c4_c.jpg

Answer

Enter your answer here.

Figure 3: A screenshot of our developed web tool to collect answers from anonymous learners. Note that all instructions are originally written in Japanese, but we translate them into English in this paper.

Correctness Edit. For this type of annotation, the experts corrected grammatical and semantic errors⁴ in the

⁴Note that “semantic error” means mismatching of a sentence and the corresponding image.

learners’ sentences. We asked them to edit the sentences while maintaining the original syntax as much as possible; for example, in Figure 4(d), the incorrect word “month” was corrected to “horse.” The second line in Table 1 lists the statistics of these annotations. There are 7 fewer sentences than in the first line because those sentences were impossible to correct without rewriting them from scratch.

Engagement Edit. We asked the experts to add further information about the images (e.g., their attributes or locations) to the learner sentences to encourage the learners to use richer expressions. For example, in Figure 4(a), the phrase “leaving a white trail with two other jets” was added along with two correctness edits. The third line in Table 1 lists the statistics of these annotations, which nearly doubled the average sentence length.

4. Data Analysis

We analyzed 100 randomly sampled learner sentences and found that 82 of them required corrections. Among these 82 sentences, only 18 were categorized as “correctable without an image,” while 64 were categorized as “uncorrectable without an image.” Although we applied a state-of-the-art GEC model, GEC-ToR (Omelianchuk et al., 2020), to these 64 sentences, almost all of them were not corrected appropriately (32 were revised by GEC-ToR, but only three of the revi-

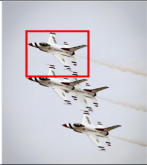
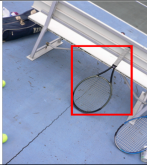


	(a)	(b)	(c)	(d)
				
Learner Sentence	That air is flying of sky.	A tennis racket is black clour on the bench.	A old man is having wain grass.	A month is lonely
Correctness Edit	A jet is flying in the sky.	The tennis racket is black and by the bench.	An old man is holding a wine glass.	A horse is lonely.
Engagement Edit	A jet is flying in the sky leaving a white trail with two other jets.	The tennis racket is black leaned on the bench.	An old man is holding a wine glass at an outdoor dinner table.	A horse is lonely, separated from other horses and sheep in a ranch.

Figure 4: Two types of corrections by experts: Correctness Edit and Engagement Edit.

Factor	Freq.
Unknown word due to misspelling	30
Inappropriate word due to misspelling	16
Inappropriate word choice	9
Wrong number of object(s)	9
Inappropriate phrase	4
Wrong relation of objects	4

Table 2: Frequent factors making learner sentences be uncorrectable without image. Each of them can include multiple factors.

sions were correct). We also found a case in which GECToR wrongly revised a correct sentence, “The bird is swimming on the water,” to “The bird is swimming in the water,” which was incorrect because the bird in the image was not underwater but floating. Such cases could be avoided through correction with image information, which demonstrates the significance of automatic assessment of image description for language learners.

We analyzed the sentences that were uncorrectable without an image in greater detail, and Table 2 lists frequent factors that caused this problem. Here, we explain each of these factors:

1. **Unknown word due to misspelling** means that the learner misspelled a word to produce a word that does not exist in English (e.g., “animaru” for “animal”).
2. **Inappropriate word due to misspelling** means that the learner produced a word that matched an English word besides the intended word and was inappropriate to describe the object, act, or situation in the image (e.g., “There is a bench on glass,” but the bench in the image is on grass).
3. **Inappropriate word choice** means that the learner chose an inappropriate word to describe the object, act, or situation in the image (e.g., “A cup which is in front of ...,” but the object in the image is a glass).

4. **Wrong number of object(s)** means that the number of objects did not match the image (e.g., “banana” and “bananas”).
5. **Inappropriate phrase** means that the learner chose a phrase that did not fit the image (e.g., “has air” for “is empty,” “is put in shadow site” for “is put in the shadow”).
6. **Wrong relation of objects** means that a relation of objects did not match the image (e.g., “The peech that is in many floots,”⁵ but the peach appears with other fruits in the image).

Note that “Inappropriate word due to misspelling” is almost the same as “Inappropriate word choice.” We did not ask the learners what they actually intended to express; the classification of “inappropriate word choice” or “inappropriate word due to misspelling” was based on our judgments. Note also that the dataset construction and analysis are ongoing: the factor set and classification criteria may change in a release version.

5. Error Correction Model for Image Description

Based on the LLID dataset, we propose a novel task of error correction of image description for language learners. Given an image I and a learner sentence L , the goal of this task is to output a corrected sentence C that is grammatically correct and describes the image contents as the learner originally intended.

Then we develop a baseline error collection model. Learners are expected to express specific objects in the image using its surrounding objects. To correct the learner sentence accurately, it is essential to incorporate object-wise representations into the text generation model effectively.

To achieve this, this paper introduces visual attention (Anderson et al., 2018; Xu et al., 2015) into a standard encoder-decoder architecture (Figure 5). The model

⁵This example also has misspellings of “peach” and “fruits.”

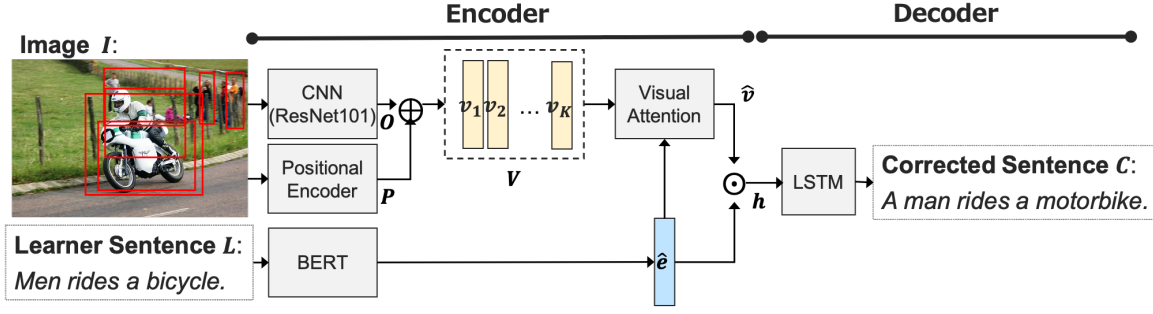


Figure 5: Overview of our error correction model.

first converts objects and a learner sentence into representations with image and sentence encoders. Then the visual attention softly calculates attention weights of objects with respect to the sentence representation and outputs attention vector using them. Based on the attention vector, the LSTM-based decoder generates a corrected sentence.

5.1. Encoder

The input of our task has two components: an image I and learner sentence L . We develop a suitable encoder for each component: image encoder and sentence encoder.

Image Encoder. The image encoder converts objects in an image I into object-wise representations. Because learners describe outside the given frame too, we need to be aware of all other objects. To detect objects from image I , we employ Faster-RCNN (Ren et al., 2015) that is one of the state-of-the-art object detection model and extract object information in the forms of bounding box coordinates. Then we input them to ResNet-101 (He et al., 2016) to obtain appearance representations of objects $O = (o_1, \dots, o_k, \dots, o_K) \in \mathbb{R}^{d_v \times K}$. In addition to the appearance, the positions of objects are also important for the model to consider the relative locations of objects. To achieve this, we use a linear layer to expand 4-dimensional bounding box coordinates to d_e dimensional position features $P = (p_1, \dots, p_k, \dots, p_K) \in \mathbb{R}^{d_e \times K}$. The object-wise representations $V = (v_1, \dots, v_k, \dots, v_K) \in \mathbb{R}^{d_e \times K}$ are computed as an element-wise addition of O and P as:

$$\begin{aligned} V &= (v_1, \dots, v_K) \\ &= (W_o o_1 + p_1, \dots, W_o o_K + p_K), \end{aligned}$$

where $W_o \in \mathbb{R}^{d_e \times d_v}$ are learnable. We call this calculation **Bottom-up**. On the other hand, we can also consider using the image feature $v' \in \mathbb{R}^{d_e}$ of ResNet-152 extracted from the entire image as the image encoder output. We call this alternative calculation **Global** to discriminate it from **Bottom-up**⁶.

⁶Global also use $W_o \in \mathbb{R}^{d_e \times d_v}$ to reduce the dimension of the ResNet-152 output from d_v to d_e .

Sentence Encoder. The sentence encoder converts a learner sentence L into a sentence representation $\hat{e} \in \mathbb{R}^{d_e}$. In our work, we use a pre-trained BERT (Devlin et al., 2019) as the sentence encoder. L is first split into a sequence of words using the WordPiece Tokenizer and further encoded into representations $(e_1, \dots, e_n, \dots, e_N) \in \mathbb{R}^{d_e \times N}$ by BERT (N is the length of the tokenized word sequence). We finally compute an average vector of them as a sentence representation \hat{e} as:

$$\hat{e} = \frac{1}{N} \sum_{i=1}^N e_i. \quad (1)$$

Visual Attention. Based on the encoded vectors of objects V and sentence \hat{e} , the visual attention calculates an attention vector \hat{v} , which softly attends objects with respect to the sentence. Following Li et al. (2018), we compute \hat{v} as:

$$\hat{v} = \sum_{k=1}^K \alpha_k v_k. \quad (2)$$

Here the attention weight α_k is computed as:

$$\alpha_k = \frac{\exp(\tau_k)}{\sum_{j=1}^K \exp(\tau_j)}, \quad (3)$$

$$\tau_k = W_{ve} (\text{ReLU}(W_v v_k) \odot \text{ReLU}(W_e \hat{e})), \quad (4)$$

where \odot is a Hadamard product, $W_v \in \mathbb{R}^{d_e \times d_e}$, $W_e \in \mathbb{R}^{d_e \times d_e}$, and $W_{ve} \in \mathbb{R}^{1 \times d_e}$ are learnable⁷. Note that if the image encoder calculates the output with the Global option, the calculation of the visual attention is skipped, and v' is used as \hat{v} in the following computations.

5.2. Decoder

Based on a pair of the sentence and attention vectors (\hat{e}, \hat{v}) , the LSTM-based decoder generates a corrected sentence. The LSTM hidden state h is initialized based on a joint representation of (\hat{e}, \hat{v}) as:

$$h = \text{ReLU}(W_f \hat{v}) \odot \text{ReLU}(W_g \hat{e}), \quad (5)$$

⁷Note that we omit the bias terms in the equations.

Model	Image Features	Synthetic Dataset	ERRANT			GLEU
			Prec.	Rec.	$F_{0.5}$	
GECToR	-	✓*	0.367	0.116	0.256	0.268
L-C	-	✓	0.105	0.126	0.108	0.226
LI-C (w/o pre-training)	Bottom-up		0.137	0.183	0.144	0.176
LI-C	Global	✓	0.147	0.168	0.151	0.263
LI-C	Bottom-up	✓	0.165	0.190	0.170	0.284

Table 3: Quantitative evaluation result*Note that GECToR is pre-trained on the original synthetic dataset, which is different from our prepared synthetic dataset.

where $W_f, W_g \in \mathbb{R}^{d_e \times d_e}$ are learnable. Based on the initialized state h , the LSTM decoder generates a sequence of words until outputting a special “[END]” symbol.

5.3. Loss Function

Given an image I and a learner sentence L as input, and a corrected sentence C as an expected output, we consider minimizing the following negative log-likelihood \mathcal{L} :

$$\mathcal{L} = - \sum_{\mathcal{D}_{train}} \log p(C|h; \theta), \quad (6)$$

where \mathcal{D}_{train} indicates the training dataset and θ indicates all of the trainable parameters in both the encoder and decoder.

6. Baseline Result

We evaluate the baseline error correction models on the constructed LLID dataset from two criteria: quantitative evaluation using traditional GEC metrics and qualitative analysis. Note that we use “correctness edit” as the target sentence C .

6.1. Experimental Setup

Hyper-parameter Settings. We train our model using Adam optimizer (Kingma and Ba, 2015). The mini-batch size is set to 64 for pre-training on synthetic errors and 8 for fine-tuning on the LLID dataset. The dimension of d_v and d_e is set to be 2,048 and 768, which are the same dimension size to the output vectors of ResNet and BERT, respectively.

Pre-training on Synthetic Dataset. Because the dataset size is limited, the training loss does not fully converge only on the LLID dataset. To tackle this problem, we first generate synthetic dataset on the MS-COCO captions to pre-train the model and then fine-tune it on the LLID dataset. Specifically, to generate the synthetic dataset, we follow the existing GEC approach (Awasthi et al., 2019) and collect 40,186 and 201,059 image/caption pairs and 5,000/25,014 image/caption pairs for training and validation, respectively⁸.

⁸The ratio of image/caption pairs is not 1:1 but 1:5 on average.

Metrics. Because this work is the first attempt to tackle the proposed task of the LLID dataset, there are no metrics to evaluate the model; thus, we utilize two commonly-used GEC metrics in our experiments, ERRANT scorer (Bryant et al., 2017) and GLEU (Napoles et al., 2015). Note that these metrics are used for GEC tasks and do not evaluate error correction referring to the image. One of our future work is to propose a new metrics of the model’s performance on such errors.

Models. We compare our baseline model with pre-trained GECToR. In addition, we conduct an ablation study to investigate the effectiveness of image information and pre-training on synthetic dataset as following:

- **L-C:** This model generates a corrected sentence from a learner sentence without an image. We consider this method to investigate how much a model can correct a sentence without an image.
- **LI-C:** This model also is equipped with the image encoder. As we described in Section 5.1, this model has two options for calculating the image encoder output: Global and Bottom-up.
- **LI-C (w/o pre-training):** This model is trained only on the LLID dataset. We consider this model to investigate whether pre-training is effective for our task.

6.2. Quantitative Evaluation

Table 3 shows precision, recall, and $F_{0.5}$ calculated with ERRANT scorer and GLEU scores for the models.

Comparison of GECToR and LI-C. Compared with GECToR, the LI-C model achieves better performance on Recall on ERRANT and GLEU, but worse on Precision and $F_{0.5}$ on ERRANT scores. The main reason of poor Precision is that the proposed LI-C changes a learner sentence destructively while GECToR is aware of minimal edit for it. To remedy this problem, introducing a copy mechanism (Zhao et al., 2019) is one of the approaches, which enables the model to be more aware of a learner sentence.

Global or Bottom-up? The bottom-up encoding approach achieves a better performance than the global one on all of the metrics. This indicates that object-wise representations are useful for correcting a learner sentence effectively.


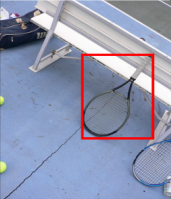


(a)		Learner Sentence	That air is flying of sky.
		Ground Truth	A jet is flying in the sky.
		GECToR	That air is flying of sky.
		LI-C (Bottom-up)	A fighter is flying in the sky.
(b)		Learner Sentence	A tennis racket is black clour on the bench.
		Ground Truth	The tennis racket is black and by the bench.
		GECToR	A tennis racket is black clour on the bench.
		LI-C (Bottom-up)	A black tennis racket is by the bench.
(c)		Learner Sentence	A old man is having wain grass.
		Ground Truth	An old man is holding a wine glass .
		GECToR	An old man is having wain grass.
		LI-C (Bottom-up)	A man is having a conversation .
(d)		Learner Sentence	A month is lonely
		Ground Truth	A horse is lonely.
		GECToR	A month is lonely
		LI-C (Bottom-up)	This is a very big .

Figure 6: Examples of generated corrected sentences. We demonstrate the outputs of GECToR and LI-C with learner sentences and ground truth. Note that red words indicate the difference from the learner sentences.

Comparison of L-C, LI-C (w/o pre-training), and LI-C. The proposed LI-C models outperform the L-C model on all of the scores, indicating the effectiveness of an image information. Moreover, we can observe that the LI-C model pre-trained on the synthetic dataset outperforms that without pre-training. This indicates the effectiveness of pre-training using a large-scale synthetic dataset. This is the same tendency to the traditional GEC tasks (Kiyono et al., 2019).

6.3. Qualitative Analysis

Figure 6 shows examples of generated corrected sentences.

Insights. Although GECToR does not utilize image information, the LI-C model can revise image-related learner’s errors (e.g., “air” to “fighter” in (a) and “black” is inserted in (b)). Moreover, the LI-C model can reflect the relation of objects on the generated sentence (e.g., “by the bench” is successfully inserted in (b)). This is a clear advantage of our LI-C model.

Limitation. We observe several failure cases, such as (c) and (d). In general, when the model fails to generate corrected sentences, it tends to ignore learner sentences. For example, in (c), a irrelevant word “conversation” is inserted instead of a “wine glass.” In (d), “horse” is deleted in the model’s output. One of the solutions to ease this problem is introducing a copy

mechanism, which encourage the model to generate corrected sentences by attending the learner sentences.

7. Conclusion

In this paper, we constructed a novel LLID dataset to support an automatic assessment system for language learners on the task of image description. The LLID dataset consists of images with descriptions written by learners and assessment annotations by English experts. By using this dataset, we proposed a novel task of automatic error correction for image description. We developed a baseline model for this task, which yields a corrected sentence by effectively using multimodal information consisting of a learner sentence and an image. Our experimental results showed that the developed model achieved better performance in terms of traditional GEC metrics than a model without images. Moreover, we confirmed that the model could revise errors that could not be revised without an image.

8. Acknowledgements

We thank everyone involved for their help with conducting the image description exercises. This work was supported by JSPS KAKENHI Grant Number 19K12119.

9. Bibliographical References

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van

- Den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proc. of CVPR*, pages 3674–3683.
- Awasthi, A., Sarawagi, S., Goyal, R., Ghosh, S., and Piratla, V. (2019). Parallel iterative edit models for local sequence transduction. In *Proc. of EMNLP-IJCNLP*, pages 4259–4269.
- Bryant, C., Felice, M., and Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proc. of ACL*, pages 793–805.
- Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *Proc. of BEA*, pages 52–75.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.
- Duan, R., Kawahara, T., Dantsuji, M., and Nanjo, H. (2020). Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis. *IEEE/ACM TASLP*, 28:391–401.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. of ICLR*.
- Kiyono, S., Suzuki, J., Mita, M., Mizumoto, T., and Inui, K. (2019). An empirical study of incorporating pseudo data into grammatical error correction. In *Proc. of EMNLP-IJCNLP*, pages 1236–1242.
- Li, Q., Tao, Q., Joty, S., Cai, J., and Luo, J. (2018). VQA-E: Explaining, elaborating, and enhancing your answers for visual questions. In *Proc. of ECCV*, pages 552–567.
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. of ECCV*.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft COCO: Common Objects in Context. In *Proc. of ECCV*, pages 740–755.
- Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proc. of CVPR*, pages 11–20.
- Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (2015). Ground truth for grammatical error correction metrics. In *Proc. of ACL-ICJNLP*, pages 588–593.
- Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Santoro, R. H., and Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proc. of CoNLL*, pages 1–14.
- Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanyski, O. (2020). GECToR – grammatical error correction: Tag, Not Rewrite. In *Proc. of BEA*, pages 163–170.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of NeurIPS*, pages 91–99.
- Settles, B., Brust, C., Gustafson, E., Hagiwara, M., and Madnani, N. (2018). Second language acquisition modeling. In *Proc. of BEA*. ACL.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proc. of CVPR*, pages 3156–3164.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of ICML*, pages 2048–2057.
- Zhao, W., Wang, L., Shen, K., Jia, R., and Liu, J. (2019). Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proc. of NAACL*, pages 156–165.