TMEKU System for the WAT2021 Multimodal Translation Task

Yuting Zhao¹, Mamoru Komachi¹, Tomoyuki Kajiwara², Chenhui Chu³

¹Tokyo Metropolitan University

²Ehime University

³Kyoto University

zhao-yuting@ed.tmu.ac.jp, komachi@tmu.ac.jp

kajiwara@cs.ehime-u.ac.jp, chu@i.kyoto-u.ac.jp

Abstract

We introduce our TMEKU¹ system submitted to the English-Japanese Multimodal Translation Task for WAT 2021. We participated in the Flickr30kEnt-JP task and Ambiguous MSCOCO Multimodal task under the constrained condition using only the officially provided datasets. Our proposed system employs soft alignment of word-region for multimodal neural machine translation (MNMT). The experimental results evaluated on the BLEU metric provided by the WAT 2021 evaluation site show that the TMEKU system has achieved the best performance among all the participated systems. Further analysis of the case study demonstrates that leveraging wordregion alignment between the textual and visual modalities is the key to performance enhancement in our TMEKU system, which leads to better visual information use.

1 Introduction

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) has achieved state-of-the-art translation performance. However, there remain numerous situations where textual context alone is insufficient for correct translation, such as in the presence of ambiguous words and grammatical gender. Therefore, researchers in this field have established multimodal neural machine translation (MNMT) tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018), which translates sentences paired with images into a target language.

Due to the lack of multimodal datasets, multimodal tasks on the English \rightarrow Japanese (En \rightarrow Ja) language pair have not been paid attention to. Since the year 2020, as the multimodal dataset on the En \rightarrow Ja language pair has been made publicly available, the multimodal machine translation (MMT) tasks on the En \rightarrow Ja were held at the WAT 2020 (Nakazawa et al., 2020) for the first time. Some studies (Tamura et al., 2020) have started to focus on incorporating multimodal contents, particularly images, to improve the translation performance on the En \rightarrow Ja task.

In this study, we apply our system (Zhao et al., 2021) for the MMT task on the En \rightarrow Ja language pair, which is called TMEKU system. This system is designed to translate a source word into a target word, focusing on a relevant image region. To guide the model to translate certain words based on certain image regions, explicit alignment over source words and image regions is needed. We propose to generate soft alignment of word-region based on cosine similarity between source words and visual concepts. While encoding, textual and visual modalities are represented interactively by leveraging the word-region alignment, which is associating image regions with respective source words.

The contributions of this study are as follows:

- 1. Our TMEKU system outperforms baselines and achieves the first place evaluated by BLEU metric among all the submitted systems in the multimodal translation task of WAT 2021^2 (Nakazawa et al., 2021) on the En \rightarrow Ja.
- 2. Further analysis demonstrates that our TMEKU system utilizes visual information effectively by relating the textual to visual information.

¹TMEKU is the abbreviation of the combination of the Tokyo Metropolitan University, the Ehime University and the Kyoto University.

²https://lotus.kuee.kyoto-u.ac.jp/WAT/ WAT2021/



Figure 1: The soft alignment of word-region.

2 TMEKU System

2.1 Word-Region Alignment

As shown in Figure 1, we propose to create an alignment between semantically relevant source words and image regions.

For the regions, we follow Anderson et al. (2018) in detecting object-level image regions from each image, which are denoted by bounding boxes on the figure. In particular, each bounding box is detected along with a visual concept consisting of an attribute class followed by an object class instead of only the object class. We take these visual concepts to represent the image regions. We set each image labeled with 36 visual concepts of image regions, which are space-separated phrases. For the words, we lowercase and tokenize the source English sentences via the Moses toolkit.³

The soft alignment is a similarity matrix filled with the cosine similarity between source words and visual concepts. To avoid unknown words, we convert the words and concepts into subword units using the byte pair encoding (BPE) model (Sennrich et al., 2016). Subsequently, we utilize fastText (Bojanowski et al., 2017) to learn subword embeddings. We use a pre-trained model⁴ containing two million word vectors trained with subword information on Common Crawl (600B tokens). The source subword embeddings can be generated directly, whereas the generation of visual concept embeddings should take an average of the embeddings of all constituent subwords because they are phrases. As shown in Figure 1, source subwords are represented by $W = {\mathbf{w_1}, \mathbf{w_2}, \mathbf{w_3}, \cdots, \mathbf{w_n}},$ and the visual concepts are represented by C = $\{c_1, c_2, c_3, \cdots, c_{36}\}$. These embeddings provide a mapping function from a subword to a 300-dim vector, where semantically similar subwords are

⁴https://fasttext.cc/docs/en/ english-vectors.html embedded close to each other. Finally, we calculate a cosine similarity matrix of the word-region as a soft alignment $A_{\rm soft}$.

2.2 Encoder

2.2.1 Representing Textual Input

In Figure 2, the textual encoder is a bi-directional RNN. Given a source sentence of *n* source words, the encoder generates the forward annotation vectors $(\overrightarrow{\mathbf{h}_1}, \overrightarrow{\mathbf{h}_2}, \overrightarrow{\mathbf{h}_3}, \cdots, \overrightarrow{\mathbf{h}_n})$, and backward annotation vectors $(\overrightarrow{\mathbf{h}_1}, \overrightarrow{\mathbf{h}_2}, \overrightarrow{\mathbf{h}_3}, \cdots, \overrightarrow{\mathbf{h}_n})$. By concatenating the forward and backward vectors $\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}]$, all words are denoted as $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n)$.

2.2.2 Representing Visual Input

We follow Anderson et al. (2018) in extracting the region-of-interest (RoI) features of detected image regions in each image. There are 36 object-level image region features, each of which is represented as a 2,048-dim vector \mathbf{r} , and all features in an image are denoted as $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \cdots, \mathbf{r}_{36})$.

2.2.3 Representations with Word-Region Alignment

As shown in Figure 2. represent we textual annotation source words of n $\mathbf{A}^{\mathrm{txt}}$ $(\mathbf{a}_1^{\mathrm{txt}}, \mathbf{a}_2^{\mathrm{txt}}, \mathbf{a}_3^{\mathrm{txt}}, \cdots, \mathbf{a}_n^{\mathrm{txt}}),$ as annotation of 36 regions as and visual $A^{img} = (\mathbf{a}_1^{img}, \mathbf{a}_2^{img}, \mathbf{a}_3^{img}, \cdots, \mathbf{a}_{36}^{img}).$

We represented the visual annotation A^{img} by concatenating R with the aligned textual features H_{align} and the textual annotation A^{txt} using textual input representation H directly.

The calculation of the $\mathrm{A}^{\mathrm{img}}$ is computed as follows:

$$\begin{aligned} \mathbf{A}^{\text{img}} &= \text{CONCAT}(\mathbf{R}, \mathbf{H}_{\text{align}}) \\ \mathbf{H}_{\text{align}} &= \frac{\mathbf{A}_{\text{soft}}^{\text{T}} \cdot \mathbf{H}}{|\mathbf{H}|} \end{aligned}$$

where the $|\mathbf{R}|$ and $|\mathbf{H}|$ represent the length of source words and the numbers of image regions: n and 36; the CONCAT is a concatenation operator.

³https://github.com/moses-smt/

mosesdecoder



Figure 2: The TMEKU system.

2.3 Decoder

To generate target word y_t at time step t, a hidden state proposal $\mathbf{s}_t^{(1)}$ is computed in the first cell of deepGRU (Delbrouck and Dupont, 2018) (GRU (1)) by function $f_{gru_1}(y_{t-1}, \mathbf{s}_{t-1})$. The function considers the previously emitted target word y_{t-1} and generated hidden state \mathbf{s}_{t-1} as follows.

$$\mathbf{s}_{t}^{(1)} = (1 - \hat{\xi}_{t}) \odot \mathbf{\dot{s}}_{t} + \hat{\xi}_{t} \odot \mathbf{s}_{t-1}$$
$$\mathbf{\dot{s}}_{t} = \tanh(WE_{Y}[y_{t-1}] + \hat{\gamma}_{t} \odot (U\mathbf{s}_{t-1}))$$
$$\hat{\gamma}_{t} = \sigma(W_{\gamma}E_{Y}[y_{t-1}] + U_{\gamma}\mathbf{s}_{t-1})$$
$$\hat{\xi}_{t} = \sigma(W_{\xi}E_{Y}[y_{t-1}] + U_{\xi}\mathbf{s}_{t-1})$$

where $W_{\xi}, U_{\xi}, W_{\gamma}, U_{\gamma}, W$, and U are training parameters, and E_Y is the target word embedding.

2.3.1 Text-Attention

At time step t, the text-attention focuses on every textual annotation $\mathbf{a}_i^{\mathbf{txt}}$ in $A^{\mathbf{txt}}$ and assigns an attention weight. The textual context vector \mathbf{z}_t is generated as follows.

$$e_{t,i}^{\text{text}} = (V^{\text{text}})^{\text{T}} \tanh(U^{\text{text}} \mathbf{s}_{t}^{(1)} + W^{\text{text}} \mathbf{a}_{i}^{\text{txt}}),$$

$$\alpha_{t,i}^{\text{text}} = \operatorname{softmax}(e_{t,i}^{\text{text}})$$

$$\mathbf{z}_{t} = \sum_{i=1}^{n} \alpha_{t,i}^{\text{text}} \mathbf{a}_{i}^{\text{txt}}$$

where V^{text} , U^{text} , and W^{text} are the training parameters; $e_{t,i}^{\text{text}}$ is the attention energy; and $\alpha_{t,i}^{\text{text}}$ is the attention weight matrix.

2.3.2 Image-Attention

Similarly, the visual context vector \mathbf{c}_t is generated as follows.

$$\begin{split} e_{t,j}^{\text{img}} &= (V^{\text{img}})^{\text{T}} \tanh(U^{\text{img}} \mathbf{s}_{t}^{(1)} + W^{\text{img}} \mathbf{a}_{j}^{\text{img}}), \\ \alpha_{t,j}^{\text{img}} &= \operatorname{softmax}(e_{t,j}^{\text{img}}) \\ \mathbf{c}_{t} &= \sum_{j=1}^{36} \alpha_{t,j}^{\text{img}} \mathbf{a}_{j}^{\text{img}} \end{split}$$

where V^{img} , U^{img} , and W^{img} are the training parameters; $\alpha_{t,j}^{\text{img}}$ is a weight matrix of each $\mathbf{a}_{j}^{\text{img}}$; and $e_{t,j}^{\text{img}}$ is the attention energy.

2.3.3 DeepGRU

As shown in Figure 2, deepGRU consists of three layers of GRU cells, which are variants of the conditional gated recurrent unit (cGRU).⁵ The hidden state s_t is computed in GRU (3) as follows. Because the calculation of f_{gru_2} and f_{gru_3} are similar to function f_{gru_1} , they are not included in the paper.

$$\mathbf{s}_t = f_{\text{gru}_3}([\mathbf{c}_t, y_{t-1}], \mathbf{s}_t^{(2)})$$
$$\mathbf{s}_t^{(2)} = f_{\text{gru}_2}(\mathbf{z}_t, \mathbf{s}_t^{(1)})$$

We use a gated hyperbolic tangent activation (Teney et al., 2018) instead of tanh. This nonlinear layer implements function $f_{ght} : \mathbf{x} \in \mathbb{R}^m \to \mathbf{y} \in \mathbb{R}^n$ with parameters defined as follows.

$$\mathbf{y}' = \tanh(K\mathbf{x} + \mathbf{b})$$
$$\mathbf{g} = \sigma(K'\mathbf{x} + \mathbf{b}')$$
$$\mathbf{y} = \mathbf{y}' \odot \mathbf{g}$$

where $K, K' \in \mathbb{R}^{n \times m}$ and **b**, **b**' $\in \mathbb{R}^n$ are the training parameters.

⁵https://github.com/nyu-dl/

dl4mt-tutorial/blob/master/docs/cgru.pdf

To ensure that both representations have their own projections to compute the candidate probabilities, a textual GRU block and visual GRU block (Delbrouck and Dupont, 2018) obtained as below.

$$\begin{aligned} \mathbf{b}_{t}^{\mathbf{v}} &= \mathbf{f}_{\text{ght}}(W_{\text{b}}^{\mathbf{v}}\mathbf{s}_{t}) \\ \mathbf{b}_{t}^{\mathbf{t}} &= \mathbf{f}_{\text{ght}}(W_{\text{b}}^{\text{t}}\mathbf{s}_{t}^{(2)}) \\ y_{t} &\sim p_{t} = \text{softmax}(W_{\text{proj}}^{\text{t}}\mathbf{b}_{t}^{\mathbf{t}} + W_{\text{proj}}^{\text{v}}\mathbf{b}_{t}^{\mathbf{v}}), \end{aligned}$$

where $W_{\rm b}^{\rm v}, W_{\rm b}^{\rm t}, W_{\rm proj}^{\rm t}, W_{\rm proj}^{\rm v}$ are training parameters.

3 Experiments

3.1 Dataset

Firstly, we conducted experiments for the En \rightarrow Ja task using the official Flickr30kEnt-JP dataset (Nakayama et al., 2020), which was extended from the Flickr30k (Young et al., 2014) and Flickr30k Entities (Plummer et al., 2017) datasets, where manual Japanese translations were newly added.

For training and validation, we used the Flickr30kEnt-JP dataset⁶ for Japanese sentences, the Flickr30k Entities dataset⁷ for English sentences, and the Flickr30k dataset⁸ for images. They were sharing the same splits of training and validation data made in Flickr30k Entities. For test data, we used the officially provided data of the Flickr30kEnt-JP task, and their corresponding images were in the Flickr30k dataset.

Note that the Japanese training data size is originally 148,915 sentences, but five sentences are missing. Thus, we used 148,910 sentences for training. In summary, we used 148,910 pairs for training, 5k pairs for validation, and 1k monolingual English sentences for translating test results.

Secondly, we also conducted experiments for the En \rightarrow Ja task using the official Ambiguous MSCOCO dataset (Merritt et al., 2020),⁹ which was extended from the Ambiguous COCO captions and images,¹⁰ where the Japanese translations were newly added. It was including a validation set with 230 pairs and a test set with 231 pairs. For standard training data, the training data from the Flickr30kEnt-JP dataset was officially designated.

3.2 Preprocessing

For English sentences, we applied lowercase, punctuation normalization, and the tokenizer in the Moses Toolkit. Then we converted space-separated tokens into subword units using the BPE model with 10k merge operations. For Japanese sentences, we used MeCab¹¹ for word segmentation with the IPA dictionary. The resulting vocabulary sizes of $En \rightarrow Ja$ were 9,578 \rightarrow 22,274 tokens.

For image regions, we used Faster-RCNN (Ren et al., 2015) in Anderson et al. (2018) to detect up to 36 salient visual objects per image and extracted their corresponding 2,048-dim image region features and attribute-object combined concepts.

3.3 Settings

(i) NMT: the baseline NMT system (Bahdanau et al., 2015) is the architecture comprised a 2-layer bidirectional GRU encoder and a 2-layer cGRU decoder with attention mechanism, which only encodes the source sentence as the input.

(ii) MNMT: the baseline MNMT system without word-region alignment (Zhao et al., 2020). This architecture comprised a 2-layer bidirectional GRU encoder and a 2-layer cGRU decoder with double attentions to integrate visual and textual features.

(iii) TMEKU system: our proposed MNMT system with word-region alignment.

We conducted all experiments on Nmtpy toolkit (Caglayan et al., 2017).

3.3.1 Parameters

We ensured that the parameters were consistent in all the settings. We set the encoder and decoder hidden state to 400-dim; word embedding to 200-dim; batch size to 32; beam size to 12; text dropout to 0.3; image region dropout to 0.5; dropout of source RNN hidden states to 0.5; and blocks \mathbf{b}_t^t and \mathbf{b}_t^v to 0.5.

Specifically, the textual annotation A^{txt} was 800dim, which was consistent with H. Further, the visual annotation A^{img} was 4,096-dim by a concatenation of R and H_{align} , where R was 2,048-dim and H_{align} was 2,048-dim by a linear transformation from 800-dim.

We trained the model using stochastic gradient descent with ADAM (Kingma and Ba, 2015) and a learning rate of 0.0004. We stopped training when the BLEU (Papineni et al., 2002) score did not improve for 20 evaluations on the validation set,

⁶https://github.com/nlab-mpg/ Flickr30kEnt-JP ⁷http://bryanplummer.com/ Flickr30kEntities/ ⁸http://shannon.cs.illinois.edu/ DenotationGraph/ ⁹https://github.com/knccch/JaEnCOCO ¹⁰http://www.statmt.org/wmt17/ multimodal-task.html

¹¹https://taku910.github.io/mecab/

Model	Test	Score
Baseline NMT	46.16	
Baseline MNMT	46.33	
TMEKU System	47.02	
v.s. baseline NMT	$\uparrow 0.86$	
v.s. baseline MNMT	$\uparrow 0.69$	
Ensemble (top 10 models)	48.57	4.7225

Table 1: Flickr30kEnt-JP task: BLEU scores and human evaluation score (full score is 5) on the $En \rightarrow Ja$.

Model	Test	Score
TMEKU System	30.23	
Ensemble (8 models)	31.04	4.4825

Table 2: Ambiguous MSCOCO task: BLEU scores and human evaluation score (full score is 5) on the $En \rightarrow Ja$.

and one validation evaluation was performed after every epoch.

3.3.2 Ensembling Models

For the Flickr30kEnt-JP task on the En \rightarrow Ja, each experiment is repeated with 12 different seeds to mitigate the variance of BLEU. At last, we choose the top 10 trained models that evaluated by BLEU scores on the validation set for ensembling.

For the Ambiguous MSCOCO task on the $En \rightarrow Ja$, each experiment is repeated with 8 different seeds to mitigate the variance of BLEU and benefit from ensembling these 8 trained models for the final testing.

3.4 Evaluation

We evaluated the quality of the translation results using the official evaluation system provided by WAT 2021. We submitted the final translation results in Japanese, which was translated from the official test data in English. On the WAT 2021 evaluation site, an automatic evaluation server was prepared and the BLEU was the main metric to evaluate our submitted translation results.

3.5 Results

In Table 1, we presented the results of the baselines and our TMEKU system on the Flickr30kEnt-JP task. We compared all the results based on BLEU scores evaluated by WAT 2021 evaluation site. For instance, the TMEKU system outperformed the NMT baseline by BLEU scores of 0.86 and outperformed the MNMT baseline by BLEU scores of 0.69 on the official test set. Our TMEKU system achieved significant improvement over both the NMT and MNMT baselines. Moreover, the result of ensembling the top 10 models has achieved the first place in the ranking of this task.

We also participated in the Ambiguous MSCOCO task on the $En \rightarrow Ja$ translation using our TMEKU system. Our reported BLEU scores are shown in Table 2, and the result of ensembling 8 models has ranked the first among all the submissions in this task.

3.6 Human Evaluation

To further validate the translation performance, a human evaluation was done by the organizers.

There are two native speakers of Japanese to rate the translation results with a score of 1 to 5 (1 is the worst and 5 is the best), who are informed to focus more on semantic meaning than grammatical correctness. There are 200 randomly selected examples for evaluation on the En \rightarrow Ja language pair of Flickr30kEnt-JP task and Ambiguous MSCOCO task, respectively.

The human evaluation scores provided by the organizers are added in Table 1 and Table 2, which have achieved the best scores among the participated systems in their respective tasks.

4 Case Study

We show two cases in Figure 3, and improvement is highlighted in green.

We perform two types of visualization for each case: (1) We visualize the source-target word alignment of the text-attention. (2) We visualize the region-target alignment of the image-attention at a time step that generates a certain target word along with attending to the most heavily weighted image region feature.

In the case shown on the left, our TMEKU system translates "entering" to "entrant," but the baselines under-translate. By visualization, the textattention and image-attention assign the highest weights to the word and region that are semantically relevant at that time step of generating "entrant." This example shows that translation quality improvement is due to the simultaneous attentions of semantically related image regions and words.

In the case shown on the right, our TMEKU system correctly translates "backyard" to a com-



Figure 3: Examples for case study. The improved translation is highlighted in green.

pound noun of "arrière-cour." But the baselines mistranslates it to "cour," which means "yard" in English. Through visualization, we find that the text-attention and image-attention focus on the features that are semantically relevant at that time step. This example shows that the image region feature associated with its semantically relevant textual feature can overcome the deficiency, where the object attribute cannot be specifically represented by only the image region feature.

5 Conclusion

We presented our TMEKU system to the English \rightarrow Japanese MMT tasks for WAT 2021, which is designed to simultaneously consider relevant textual and visual features during translation. By integrating the explicit word-region alignment, the object-level regional features can be further specified with respective source textual features. This leads the two attention mechanisms to understand the semantic relationships between textual objects and visual concepts.

Experimental results show that our TMEKU system exceeded baselines by a large margin and achieved the best performance among all the participated systems. We also performed analysis of case study to demonstrate the specific improvements resulting from related modalities.

In the future, we plan to propose a more efficient integration method to make modalities interactive with each other.

Acknowledgments

This work was supported by Grant-in-Aid for Young Scientists #19K20343, JSPS. We are immensely grateful to Mr. Tosho Hirasawa, who provided coding support, and comments that significantly help the implementation of our system.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR*, abs/1409.0473.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *WMT*, pages 304–323.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.
- Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. NMTPY: A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, 109:15–28.
- Jean-Benoit Delbrouck and Stéphane Dupont. 2018. UMONS submission for WMT18 multimodal translation task. In *WMT*, pages 643–647.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the

second shared task on multimodal machine translation and multilingual image description. In *WMT*, pages 215–233.

- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*, pages 1–15.
- Andrew Merritt, Chenhui Chu, and Yuki Arase. 2020. A corpus for english-japanese multimodal neural machine translation with comparable sentences. *CoRR*, abs/2010.08725.
- Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *LREC*, pages 4204–4210.
- Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *WAT*.
- Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *WAT*, pages 1–44.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311– 318.
- Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, pages 74–93.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *WMT*, pages 543–553.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.
- Hiroto Tamura, Tosho Hirasawa, Masahiro Kaneko, and Mamoru Komachi. 2020. TMU Japanese-English multimodal machine translation system for WAT 2020. In WAT, pages 80–91.

- D. Teney, P. Anderson, X. He, and A. v. d. Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, pages 4223–4232.
- 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. *TACL*, 2:67–78.
- Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *EAMT*, pages 105–114.
- Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Neural machine translation with semantically relevant image regions. In *NLP*.