DejaVu: Disambiguation evaluation dataset for English-JApanese machine translation on VisUal information

Ayako Sato[1], **Tosho Hirasawa**[1], **Hwichan Kim**[1], **Zhousi Chen**[2]
Teruaki Oka[2], **Masato Mita**[1,3], **Mamoru Komachi**[2]
[1]Tokyo Metropolitan University, [2]Hitotsubashi University, [3]CyberAgent Inc.
{sato-ayako, kim-hwichan}@ed.tmu.ac.jp
toshosan@tmu.ac.jp, mita_masato@cyberagent.co.jp
{zhousi.chen, teruaki.oka, mamoru.komachi}@r.hit-u.ac.jp

Abstract

Multimodal machine translation (MMT) should resolve textual translation ambiguity given visual content completion. However, general MMT benchmarks are not featured in the evaluation of this capacity because caption texts are self-disambiguating and barely necessitating visual information. To address this issue, we focus on word sense disambiguation (WSD) and propose the English-Japanese WSD-oriented MMT evaluation dataset, DejaVu. For efficiency and coverage of data curation, DejaVu automatically retrieves ambiguous words and houses each in a simple caption template with images as the only disambiguating means for their correct translations. The effectiveness of DejaVu is demonstrated by comparison experiments with existing benchmarks. Evaluation with DejaVu exhibited the presence of image-based WSD capabilities in the latest vision language models. Our dataset is publicly available at the following URL [1].

1 Introduction

The fusion of natural language processing and computer vision has attracted much attention. As an advance of such fusion, multimodal machine translation (MMT) resorts to visual information for ambiguous textual concepts, whereas text-only machine translation (MT) fails by pure chance. For instance, in Figure 1, the images provide meaningful clues to disambiguate *"seal"* and determine the correct translations in Japanese. This completion is expected to yield an effect of resolving ambiguities in word-sense, syntax, and grammaticality.

The de-facto benchmarks for MMT are constructed by translating English captions from the Flickr30k dataset (Young et al., 2014) into German (Elliott et al., 2016), French (Elliott et al., 2017), Czech (Barrault et al., 2018), and Japanese (Nakayama et al., 2020). Since the English captions describe the images in detail with no ambiguity, most of them do not require completion with visual information for generating precise translations (Frank et al., 2018). About 1-2% (Futeral et al., 2023) or 5-6% (Frank et al., 2018) of such image-demand cases have been reported. Therefore, Flickr30k limits the depth of evaluation on the disambiguation capability of MMT models.

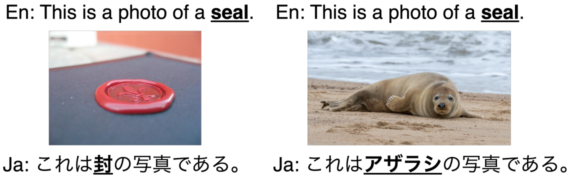

En: This is a photo of a **seal**.
En: This is a photo of a **seal**.
Ja: これは**封**の写真である。
Ja: これは**アザラシ**の写真である。

Figure 1: Visual content resolves lexical ambiguity of word **seal** for English-to-Japanese translation in DejaVu.

For a precise evaluation of the MMT system's ability to utilize multimodal information, Futeral et al. (2023) proposed the disambiguation-oriented English-French dataset CoMMuTE. When translating English sentences in CoMMuTE, the textual context is insufficient for disambiguation, so the correct translation can be achieved by referring to the corresponding images. A similar evaluation dataset for English-Japanese translation MMT systems is desirable. However, CoMMuTE has a relatively complex methodology that incorporates various caption formats. On one hand, CoMMuTE is expensive to construct, as they manually collected 29 ambiguous English sentences from Bawden et al. (2018) and self-created additional 126 sentences. On the other hand, the effect of these realistic expressions varies from instance to instance, which introduces instability during lexical-based evaluation irrelevant to WSD.

We construct a congruent dataset for English-to-Japanese MMT evaluation and title it DejaVu. It features in addressing CoMMuTE's issues of

[1] https://github.com/tmu-nlp/DejaVu

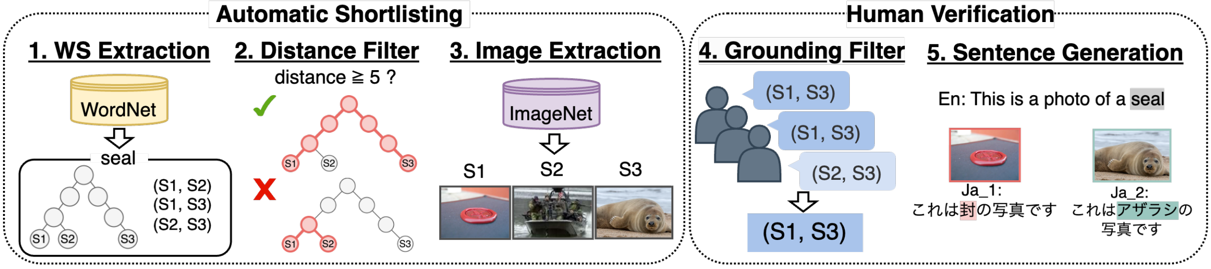


Figure 2: Overview of dataset construction. S1, S2, and S3 denote three different senses.

construction complexity and evaluation instability of lexical-based metrics. Concretely, we propose an automatic method of extracting ambiguous English words from WordNet (Fellbaum, 1998) in order to reduce construction costs, increase ambiguous word coverage, and expand data size. Further, we adopt a few templates to unify the caption format for a precise evaluation focused on the target word. This method can be easily applied to other language pairs.

We conduct experiments to assess how well the latest vision language models (VLMs) are able to utilize multimodal information as MMT systems. Assuming that those models already perform reasonably well on vision language tasks, we use them to reflect the difference between Flickr30k and DejaVu. As a result, Flickr30k fails to stimulate and evaluate those models' multimodal capacity for WSD, whereas DejaVu succeeds. In other words, DejaVu's methodology is effective and suitable for MMT evaluation.

2 Construction of DejaVu

2.1 Dataset Design

The scheme of the input/output of the DejaVu dataset is as follows: each instance consists of an English sentence, two Japanese translations, and an image corresponding to each translation. However, to address the limitations in CoMMuTE (as described in §1), the following four requirements were first established: (1) English captions contain words whose senses are ambiguous when translated from English to Japanese. (2) Word-senses can be distinguished by visual information. (3) English captions do not provide a conducive context for WSD. (4) Focusing on ambiguous target words in evaluation.

To satisfy requirements (1) and (2), we automatically collect ambiguous words and corresponding images from WordNet (Fellbaum, 1998) and ImageNet (Russakovsky et al., 2015), respectively, and those candidates are filtered by human annotators over multiple steps. In WordNet, English words are classified into groups of senses and their relationships to other groups described in tree structures. ImageNet is a large dataset of color images, and supervised labels are assigned to the images based on the tree structure of WordNet. Then, to satisfy requirements (3) and (4), we insert target words into simple, unified caption templates to generate sentences. Figure 2 shows an overall schematic description of the data construction process.

Since this method can be used to automatically extract English words with ambiguous senses and their sense pair sets from WordNet (§2.2), it is worth noting that it is possible to efficiently expand the dataset for from-English language pairs other than English-Japanese (En-Ja). Human verification by native annotators is necessary to improve the quality of the dataset (§2.3).

2.2 Automatic Shortlisting

During this phase, we extract nouns and their word-sense sets from WordNet and retrieve corresponding images from ImageNet. Although WordNet contains many specialized nouns, such as plants and animals, we aim to select words that are general enough to identify object names from images.

Step 1: Word-senses Extraction from WordNet We extract polysemous nouns and tree-structured word-senses from WordNet according to the following conditions. (1) Length less than 10 characters (to extract general words). (2) Belonging to a physical entity (to extract word-senses that can be represented by images). Then, we create word-sense pairs from the extracted word-senses.

Step 2: Distance Filter The distance between word-senses is defined as the number of edges connecting two sense nodes. We exclude word-sense pairs with a distance of less than 5 [2] (to exclude

[2] We set this parameter based on preliminary experiments.

No.	English Source Sentence	Japanese Translation References
1	This is a photo of a/an/the [ ] .	これは [ ] の写真 {である / だ / です} 。
2	It must be the [ ] .	それは [ ] {に違いない / です }。
3	Why is the [ ] here?	なぜここに [ ] があるん{だ / ですか} ?
4	I don't give a damn about the [ ] .	私は [ ] {のことはどうでもいい / に興味はありません}。
5	Can you not see the [ ] ?	[ ] が見え{ないのか / ませんか} ?
6	Look at the [ ] !	[ ] を{見て/ 見てください} !

Table 1: A full list of caption templates used in DejaVu. The target word is inserted at "[ ]". To mitigate the effect of non-essential perturbations in translations (e.g., different endings of Japanese references), we created two or three reference sentences mentioned in the "{ }" bracket and reported the average of the scores for each as the result for the template.

pairs in which the word-sense differences are so obscure that they cannot be distinguished by referring to the images). For each word, we sort word-sense pairs in descending order of distance.

Step 3: Image Extraction from ImageNet We retrieve the first image corresponding to each word-sense from ImageNet. The pairs where either node has no corresponding images are dropped.

2.3 Human Verification

After automatic shortlisting, we obtain 725 words where the average number of word-sense pairs of each word is 2.07. During this phase, we manually select appropriate pairs from the automatically extracted pairs. Besides, if the images corresponding to the selected pairs are inappropriate, we replace the images. The annotations were conducted by three people, all native Japanese speakers and master's students in Computer Science. They select word-sense pairs from the list in the same order.

Step 4: Grounding Filter We check the word-sense pairs and select the best pairs in that their word-senses are general and can be linked to different visual entities. If there is no appropriate pair, the target word is excluded. Table 6 in Appendix A shows examples of inappropriate word-sense pairs that should be excluded. Among the pairs selected by the annotators, 235 pairs were selected by one person, 81 by two persons, and 26 by three persons. If more than one pair is selected for each word, the word-sense pair selected by the most annotators is finally selected[3]. The selected words are translated into two senses in Japanese by the annotators.

Words	Images	Sentences	Average Distance
250	500	3,000	9.38

Table 2: Statistics of DejaVu. Average distance indicates the average of word-sense distances in WordNet.

Step 5: Sentence Generation We create sentences by inserting the target words into the caption templates. In addition to the intuitive template 1, five others (templates 2-6 in Table 1) were selected from CoMMuTe with our manual Japanese translations in order to create more realistic scenarios. Dedicated to image-based WSD, all templates should provide limited or no context for disambiguating the target words. Otherwise, it will be vague to conclude the contribution from images or captions. We select six templates that satisfy this standard for DejaVu. Table 2 shows the statistics of DejaVu.

To ensure that the images properly represent the corresponding word-sense and have enough quality for feature extraction, we also ask annotators to subjectively evaluate whether the images are appropriate or not. The 123 images that were judged inappropriate by one or more people (i.e., remarkably low resolution, incorrect word-sense label) are replaced with alternative images retrieved from Flickr under the CC BY license.

3 Experiment

In this experiment, we confirm the suitability of DejaVu as a dataset for evaluating the ability of En-Ja MMT systems to utilize multimodal information. We compare the performance of VLMs on the Flickr30k Entities-JP (Nakayama et al., 2020) test set and DejaVu. Based on the assumption that state-of-the-art VLMs are superior in vision and language tasks (Akiba et al., 2024), we can say that the dataset is valid if WSD performance is im-

[3] To augment DejaVu, we select 53 word-sense pairs from CoMMuTE and The Word-in-Context Dataset (Pilehvar and Camacho-Collados, 2019), which are high-quality WSD datasets that were constructed manually. We finally obtain 250 pairs by combining the word-sense pairs in Step 2.

		Flickr30k		CoMMuTE			DejaVu (Ours)		
Model	Image	BLEU	COMET	BLEU	COMET	LA	BLEU	COMET	LA
EvoVLM	✗	**30.42**	**97.16**	5.35	90.93	39.00	**23.45**	**93.76**	27.47
	✓	25.37	96.96	**10.64**	**92.98**	**53.00**	23.08	93.28	**35.77**
GPT-4o	✗	**32.42**	**96.80**	29.72	92.64	40.00	32.66	93.04	30.17
	✓	31.07	96.78	**32.59**	**93.55**	**57.00**	**35.12**	**93.73**	**42.86**

Table 3: Results of the w/ image setting vs. the w/o image setting on vision language models.

		template 1			template 2			template 3			template 4			template 5			template 6		
Model	Image	B	C	L	B	C	L	B	C	L	B	C	L	B	C	L	B	C	L
EvoVLM	✗	39.3	95.3	29.0	**12.9**	**87.6**	26.4	28.3	97.0	26.6	24.9	93.5	27.6	**10.6**	**94.4**	28.2	**24.7**	**94.8**	27.0
	✓	**43.5**	**95.3**	**37.4**	11.1	87.5	**38.4**	**33.4**	**97.6**	**37.0**	**26.7**	**95.1**	**33.4**	2.3	89.8	**32.8**	21.5	94.3	**33.1**
GPT4o	✗	40.7	95.3	32.6	27.4	87.4	29.7	14.6	96.8	31.3	**16.4**	90.1	28.9	43.2	**95.4**	29.4	**53.7**	93.2	32.7
	✓	**46.8**	**95.9**	**47.5**	**31.8**	**89.2**	**45.9**	**18.7**	**97.1**	**46.5**	14.7	**90.5**	**37.9**	**45.6**	95.4	**38.9**	53.2	**94.2**	**40.2**

Table 4: Results of each caption template of DejaVu on vision language models. B denotes BLEU, C denotes COMET, and L denotes Lexical Accuracy.

proved by supplementing visual information. We perform machine translation with w/ image (MMT) and w/o image (MT) settings, and if the performance of the w/ image setting is higher than that of the w/o image setting, we consider that the visual information is complementary. In order to company DejaVu's scheme, we provide a manual translation of CoMMuTE En-Ja[4] as a comparison. The additional experiments on in-house trained MMT models are described in Appendix C.

3.1 Settings

Models We use EvoVLM (Akiba et al., 2024) and GPT-4o ("gpt-4o-2024-05-13") (OpenAI, 2024) for our experiments. The prompts used in the experiments were created based on Robinson et al. (2023), the latest work investigating ChatGPT for MT[5]. According to them, few-shot prompts offered marginal improvements, so we conducted the experiment only with the zero-shot setting. We report the averaged results over three runs.

Metrics In addition to sacreBLEU (Post, 2018) and COMET (Rei et al., 2020), we employ a metric from Lala and Specia (2018), which calculates the score as $\frac{C}{N}$, where C is the number of times the target word in the output matched the target word in the reference precisely and N is the dataset size. We refer to this metric as Lexical Accuracy (LA). LA and COMET are presented as percentages.

BLEU and COMET are general sentence-level MT metrics, whereas LA lets us focus on the target words in templates and avoid the perturbation from the context. Thus, LA is expected to properly evaluate the WSD capacity in our scheme across all templates and models.

3.2 Results

Table 3 shows BLEU, COMET, and LA for VLMs on Flickr30k En-Ja, CoMMuTE En-Ja, and DejaVu. We evaluate image-based WSD performance by comparing settings with and without images.

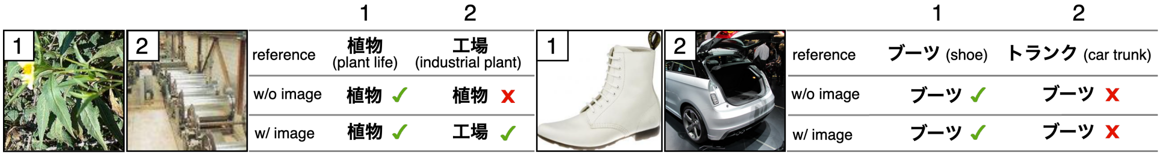
On the Flickr30k test set, we found that the without-image setting scored higher than or similar to the with-image setting. This means that while Flickr30k can be used to compare the translation performance of these models, it is not appropriate for evaluating their WSD performance.

By contrast, on CoMMuTE, the with-image setting outperforms the without-image setting, confirming that stimulating visual completion improves WSD performance. However, some examples (See Section 3.3) suggest that rich non-target words cause large oscillations, which results in significantly lower reference-based BLEU scores. That is to say, there is room for a more accurate evaluation.

On DejaVu, the performance of the settings with images in all metrics for GPT-4o and LA for EvoVLM-JP outperform that without images, respectively. The LA score is not affected by perturbations of non-target words and is dedicated to the evaluation of WSD capability, and this result reflects the intrinsic WSD capability of these VLMs.

[4] After translating the French captions into Japanese by DeepL, we manually corrected the translations by looking at the corresponding images. It will be publicly available.

[5] See Appendix B for the details of the prompts.



(a) src: This is a photo of a **plant**.

(b) src: This is a photo of a **boot**.

Figure 3: Some examples of target words in the GPT-4o outputs on DejaVu. **Bold** indicates target words.

Furthermore, DejaVu's BLEU score is higher than CoMMuTE, benefiting from unifying the templates for references.

The DejaVu results in Table 3 are the average performance for each of the six templates, and the scores for each template are shown in Table 4. The simplest caption, template 1, confirms the contribution of the image for both models in all scores. For all templates, the LA score is higher for the setting with images than for the setting without images, indicating that the WSD ability can be verified regardless of the template. However, for the other metrics, especially for templates 5 and 6, the performance is low and the setting without images is superior.

3.3 Case Study

Figure 3 shows two examples from the GPT-4o outputs on DejaVu (template 1). In the **plant** example (a), the two senses were properly discerned via the images, and the target words were correctly translated, whereas in the **boot** example (b), the word-senses were not discerned despite the visual inputs. Consistent with the results of the automatic evaluation, GPT-4o's strong image-based WSD capability is confirmed, but there is still potential for improvement. In brief, DejaVu is capable of validating image-based WSD capabilities in both quantitative and qualitative evaluations, which can be taken as a benchmark for the capacity to utilize multimodal information.

Table 5 shows some of the CoMMuTE examples from the GPT-4o outputs in Section 3.2. In the with-image setting, this shows that the target word plant is correctly translated into the two senses of "植物 *(plant life)*" and "工場 *(industrial plant)*". However, the non-target parts of the caption also change depending on the difference in input images. Despite the success of WSD in both sentences, the reference-based BLEU score, which is the de-facto evaluation metric in machine translation, is sensitive to such surface changes. To minimize the effect of such caption formatting, we use templates that simplify the non-target word parts and allow comparison of translations of only the target word.

src	So you see, they don't even own the plant.
ref	だから、彼らは植物さえも所有していない。 だから、彼らは工場さえも所有していない。
hyp	ですから、彼らはその植物を持っていない。 それで、彼らはその工場さえ所有していない。

Table 5: Some output examples of CoMMuTE En-Ja on GPT-4o. hyp is the output of the setting with images. Underline indicates target words.

We analyze the reason why the BLEU and COMET scores for the EvoVLM-JP output in templates 5 and 6 show different trends from the other templates. The output of these two templates contains looped messages that are output when the model fails to follow the instruction. We used only the base models for our VLMs experiments, and the instruction tuning data for these models probably contains a large portion of non-translation task data (or possibly none at all). Low following capability to the translation task leads to lower evaluation scores because it does not produce the expected formatted output. In addition, template 5 (Can you not see the [ ] ?) is a question sentence with negation, and template 6 (Look at the [ ] !) is an imperative sentence, which is often not included in the training data and may contain a difficult grammar for the model.

4 Related Work

In Lala and Specia (2018), the Multimodal Lexical Translation Dataset was constructed to investigate to which extent visual or textual context contributes to translation. This dataset does not focus on visual context and includes words that cannot be represented by images, making it unsuitable for evaluating the contribution of visual context in MMT. On the flip side, we construct an MMT evaluation dataset for disambiguation by visual context only.

DejaVu is synthetic data with a simple template, so we are unable to evaluate WSD capability in a real-world setting with longer sentence lengths of increased lexical and syntactic complexity. For evaluating translation performance in real-world scenarios that are not WSD-specific, one can use Flickr30k or MSCOCO (Lin et al., 2014) from the WMT multimodal shared task.

5 Conclusion

We created a WSD-oriented En-Ja MMT dataset, called DejaVu, to evaluate the capacity of MMT systems to utilize visual information. In the experiments with the latest VLMs as MMT systems, the images from the DejaVu scheme improved the scores in contrast to existing MMT benchmarks, confirming its effectiveness in assessing the contribution of visual information to the performance.

References

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2024. Evolutionary optimization of model merging recipes. *Preprint*, arXiv:2403.13187.

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 *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*.

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 *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Stella Frank, Desmond Elliott, and Lucia Specia. 2018. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24:393 – 413.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Taku Kudo. 2006. MeCab: Yet another part-of-speech and morphological analyzer. `http://taku910.github.io/mecab/`.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4204–4210, Marseille, France. European Language Resources Association.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

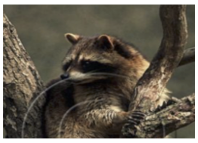
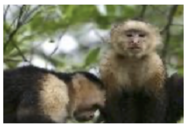
Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

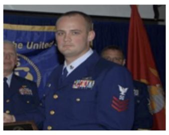
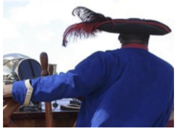
Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

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

(a) Animal, plant, and fish names are not general.

En	Ringtail (racoon)	Ringtail (monkey)
Ja	アライグマ	オマキザル
		

(b) Difficult to distinguish from visual information.

En	Captain (skipper)	Captain (lieutenant)
Ja	船長	大尉
		

(c) They are the same word in Japanese.

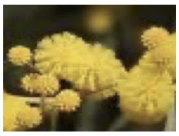
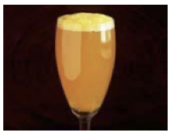
En	Mimosa (flower)	Mimosa (drink)
Ja	ミモザ	ミモザ
		

Table 6: Examples of instances excluded by human annotation and the reasons for their exclusion.

A Annotation Guideline

Table 6 shows some instances ruled out in Step 4 of Section 2.3. When the senses are too specific, a model tends to have general terms (e.g., hypernyms) as translation and the surface metrics will not catch them properly. Thus, our goal is to select word sense pairs that are general enough to identify entities from images. The annotators are instructed to exclude those candidates when any word-sense causes translating ambiguity into Japanese.

B Prompt Templates

We provide the prompt templates employed in the VLMs experiment (Section 3.2) in Table 7. Prompt templates were created based on Robinson et al. (2023). Note that ChatGPT receives images through a message apart from the text; no image appears in the prompt.

setting	prompt
w/ image	This is an English to [TGT] translation, please provide the [TGT] translation for this sentence. Do not provide any explanations or text apart from the translation. English: [src-sentence] [TGT]:
w/o image	This is an English to [TGT] translation with an image, please provide the [TGT] translation for this sentence and image. Do not provide any explanations or text apart from the translation. English: [src-sentence] [TGT]:

Table 7: Prompt templates used for w/ image and w/o image settings. In our study, [TGT] is Japanese.

C Evaluation on in-house trained models

C.1 Settings

We used DejaVu for evaluation and Flickr30k Entities-JP for both training and evaluation. Flickr30k Entities-JP has $29,000$ training data, $1,014$ validation data, and $1,000$ evaluation data. English is tokenized according to Multi30K task 1 (Elliott et al., 2016), and Japanese is word segmented by using MeCab (Kudo, 2006) (IPA dictionary). Subword segmentation is performed by using BPE (Sennrich et al., 2016).

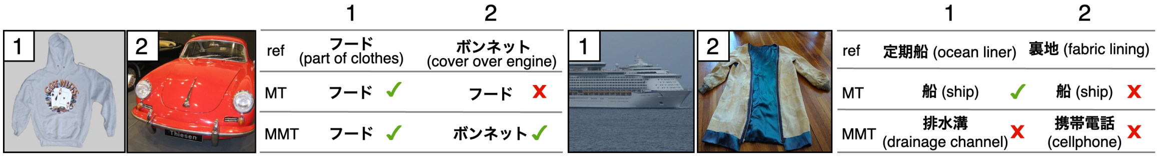
We compared in-house trained MMT models with an MT model to evaluate the contribution of images. We used Transformer-Tiny (Wu et al., 2021) as a text-based MT model. We used the Transformer-based Attentive multimodal Transformer (Attentive) (Libovický et al., 2018), Gated multimodal Transformer (Gated) (Wu et al., 2021), and Visual Translation Language Modelling (VTLM) (Caglayan et al., 2021) as MMT models. VTLM is pre-trained on the Conceptual Captions dataset. The model proposed in the previous study in which CoMMuTE was introduced requires pre-training on large amounts of caption data and we did not use it in this study due to its computational cost. We used as image features CLIP (Radford et al., 2021) based on the Vision Transformer (Dosovitskiy et al., 2021), Faster R-CNN (Ren et al., 2015), and ResNet-50 (He et al., 2016). The number of features is 1 for CLIP and ResNet-50 and 36 for Faster R-CNN.

C.2 Results

Table 8 shows the automatic evaluation scores of the existing MT and MMT models on the En-Ja MMT data. On DejaVu, the MMT model scores almost all higher than the MT model. In other words, it confirms the image-based WSD capability of the existing models.

		Flickr30k		DejaVu		
Model	ImgFeature	BLEU	COMET	BLEU	COMET	LA
Text-only Machine Translation						
Transformer	N/A	43.42	96.79	29.40	88.88	19.00
Multimodal Machine Translation						
Gated	CLIP	**43.48**	96.72	**29.68**	**93.14**	**19.60**
	ResNet	**44.12**	96.73	**30.07**	**93.44**	18.60
Attentive	CLIP	**44.48**	**96.88**	**30.43**	**93.99**	**19.60**
	R-CNN	**43.99**	**96.92**	**31.69**	**93.81**	**19.80**
VTLM	R-CNN	39.81	96.45	27.90	**94.12**	**22.00**

Table 8: Results of (M)MT models. **Bold** indicates that it outperforms the MT model.



	1	2
ref	フード (part of clothes)	ボンネット (cover over engine)
MT	フード ✓	フード ✗
MMT	フード ✓	ボンネット ✓

(a) src: This is a photo of a **hood**.

	1	2
ref	定期船 (ocean liner)	裏地 (fabric lining)
MT	船 (ship) ✓	船 (ship) ✗
MMT	排水溝 (drainage channel) ✗	携帯電話 (cellphone) ✗

(b) src: This is a photo of a **liner**.

Figure 4: Some examples of target words in the in-house trained model outputs. **Bold** indicates target words.

C.3 Case Study

We also run an in-depth analysis of the system outputs. Figure 4 shows two output examples: the MT model is Transformer-Tiny, and the MMT models are (a) VTLM (RCNN), and (b) Attentive (RCNN). In the **hood** example (a), the MT model translated both word-senses to *“フード (part of clothes)”*, whereas the MMT model was able to distinguish it from *“ボンネット (Cover over engine)”* by referring to the corresponding images. However, we found only 8 examples that the MMT model translated to the correct target words. There were also several examples in which words other than the target words were changed (e.g., insertion of the reading mark). These results suggest that an improvement in the automated evaluation score may be significantly influenced by changes in the number of tokens that are due to changes other than target words.

Although only 8 examples yielded improvement in translation quality, there were several examples in which visual information may have affected target words in the outputs (e.g., **liner** in Figure 4 (b)). Table 9 shows the number of such sentence pairs for each model. Only 3.4% of the pairs in which the translation has changed according to the image were translated correctly, that is, the existing in-house trained models utilize only modest visual information in WSD, and there is room for improvement. GPT-4o correctly translated far more words than the in-house trained model, suggesting that GPT-4o has stronger image-based WSD capability.

Model	Correct	Mislabeled	Others
Gated (CLIP)	0	0	5
Gated (ResNet)	0	3	6
Attentive (CLIP)	1	3	8
Attentive (R-CNN)	1	35	97
VTLM (R-CNN)	6	56	12
GPT-4o	116	24	1

Table 9: Number of sentence pairs in which the translation has changed according to the image. Correct is a pair in which both senses of the target word are translated correctly; Mislabeled is a pair in which at least one of the senses is translated incorrectly; Others is a pair in which the translation of the rest of the target word has changed.