English-to-Japanese Multimodal Machine Translation Based on Image-Text Matching of Lecture Videos

Ayu Teramen Takumi Ohtsuka Risa Kondo Tomoyuki Kajiwara Takashi Ninomiya
Graduate School of Science and Engineering, Ehime University, Japan
{teramen@ai., ohtsuka@ai., kondo@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp

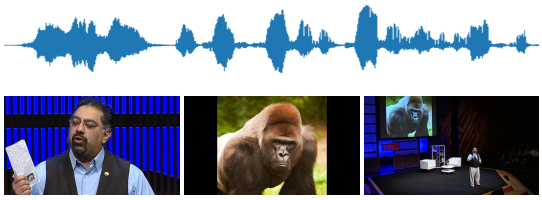
Abstract

We work on a multimodal machine translation of the audio contained in English lecture videos to generate Japanese subtitles. Image-guided multimodal machine translation is promising for error correction in speech recognition and for text disambiguation. In our situation, lecture videos provide a variety of images. Images of presentation materials can complement information not available from audio and may help improve translation quality. However, images of speakers or audiences would not directly affect the translation quality. We construct a multimodal parallel corpus with automatic speech recognition text and multiple images for a transcribed parallel corpus of lecture videos, and propose a method to select the most relevant ones from the multiple images with the speech text for improving the performance of image-guided multimodal machine translation. Experimental results on translating automatic speech recognition or transcribed English text into Japanese show the effectiveness of our method to select a relevant image.

1 Introduction

Multimodal machine translation (Sulubacak et al., 2020) is a machine translation (MT) approach that combines information from modalities other than text, such as audio and images. Since images provide visual information that is not included in audio or text, it is expected to improve translation quality by correcting errors in automatic speech recognition (ASR) or by complementing information in ambiguous text.

This study tackles the task of translating English audio or subtitles from lecture videos into Japanese. In such situations, since useful information can be obtained from the images in the presentation materials, image-guided MT can improve translation quality over text-only MT. However, some of the images derived from lecture videos are not directly related to the subtitle text, such as the image shown on the left in Figure 1, which shows only the speaker. No improvement in translation quality can be expected from such images.



ASR text (English)	when you do that here's what I can promise you're going to be the 800-pound gorilla in the forest
Transcribed text (English)	So when you do that, here's what I can promise: You're going to be the 800 pound gorilla in the forest.
Translation (Japanese)	手書きの手紙の力によって、森にいる体重 300 キ ロの ゴリラのような存在になれるでしょう。

Figure 1: An example of our multimodal parallel corpus. Our corpus includes five sets of images, audio in English, ASR sentences in English, transcribed sentences in English, and reference translations in Japanese. Three images are included, corresponding to the beginning, middle, and end of the audio.

To improve the performance of image-guided MT, we propose a method to select the image most relevant to the text among multiple images that correspond in time to the subtitle text. Additionally, to evaluate our method, we construct a multimodal parallel corpus, TAIL[1] (English-to-Japanese **T**ranslation Corpus with **A**udio and **I**mages from **L**ecture Videos), consisting of English subtitles of lecture videos and their Japanese translations. Experimental results on translating ASR or transcribed English text into Japanese subtitles show the effectiveness of our method to select a relevant image.

[1] https://github.com/EhimeNLP/TAIL

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 86–91

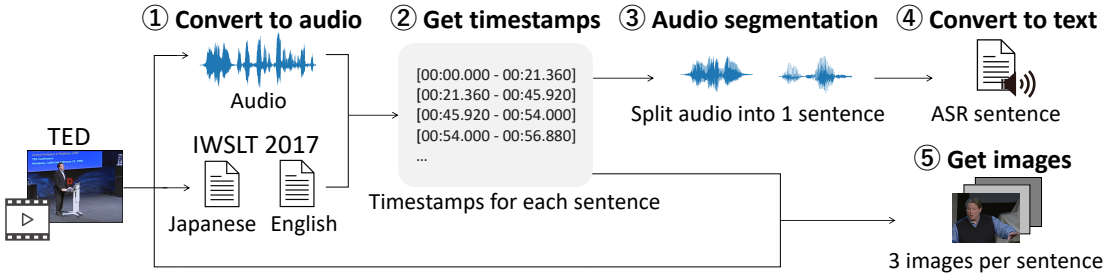


Figure 2: Overview of our corpus construction.

2 Related Work

2.1 Multimodal Parallel Corpus

Previous studies of multimodal MT have often involved adding some one modality to the text, such as MT from speech (Di Gangi et al., 2019; Wang et al., 2020; Salesky et al., 2021) or image-guided MT (Elliott et al., 2016; Parida et al., 2019; Thapliyal et al., 2022). Furthermore, we can expect further improvement in translation quality by combining three modalities of text, audio, and images. Previous studies combining the three modalities include video-guided MT, such as How-2 (Sanabria et al., 2018) and QED (Abdelali et al., 2014). However, these are limited in scope because How-2 only covers English-Portuguese language pairs and QED only covers education domain. To cover English-Japanese lecture subtitles, we need to expand the multimodal parallel corpus.

2.2 Image-guided Machine Translation

Image-guided MT (Specia et al., 2016) improves translation quality by complementing textual ambiguity with visual information derived from images. Early studies (Caglayan et al., 2016; Libovický and Helcl, 2017; Calixto and Liu, 2017) combined CNN-based visual representations with textual representations in RNN-based encoder-decoder models. In the modern approach (Li et al., 2022a), both vision and language inputs are encoded by the Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2021), integrated by selective attention, and fed to the Transformer decoder. The image-guided machine translation model, based on the powerful Vision Transformer (Dosovitskiy et al., 2021), achieves higher translation quality with images that are more relevant to the text (Yuasa et al., 2023). Therefore, in situations where multiple images are available, translation quality can be improved by selecting images that are more relevant to the text.

2.3 Vision and Language Pre-training

In image-text matching, CLIP (Radford et al., 2021) and BLIP (Li et al., 2022b), trained by multimodal contrastive learning, have achieved state-of-the-art performance. Especially, BLIP is trained in a multi-task learning manner of image-text matching and image caption generation as well as contrastive learning, which allows a single model to perform both understanding and generating on vision and language tasks.

3 TAIL Corpus

For English-to-Japanese multimodal MT of lecture subtitles, we construct a corpus consisting of five sets of images, audio in English, ASR sentences in English, transcribed sentences in English, and reference translations in Japanese for lecture videos from TED.[2] Since an English-Japanese parallel corpus consisting of transcribed sentences for TED lecture videos has been released in the IWSLT2017 competition (Cettolo et al., 2017), we annotate it with images, audio, and ASR sentences, as shown in Figure 2.

3.1 Audio Annotation

First, we annotate both audio and ASR sentences in English on top of the IWSLT2017 En-Ja corpus.

Audio Acquisition We downloaded the lecture videos in MP4 format from the URLs provided in the metadata of the IWSLT2017 En-Ja corpus. These videos are converted to audio in FLAC format with `ffmpeg` converter.[3] (Step 1 in Figure 2)

Forced Alignment For each lecture video, the transcribed English sentences from the IWSLT2017 En-Ja corpus and the audio from Step 1 are aligned by `aeneas` toolkit.[4] Here, both

[2] https://www.ted.com
[3] https://ffmpeg.org
[4] https://github.com/readbeyond/aeneas

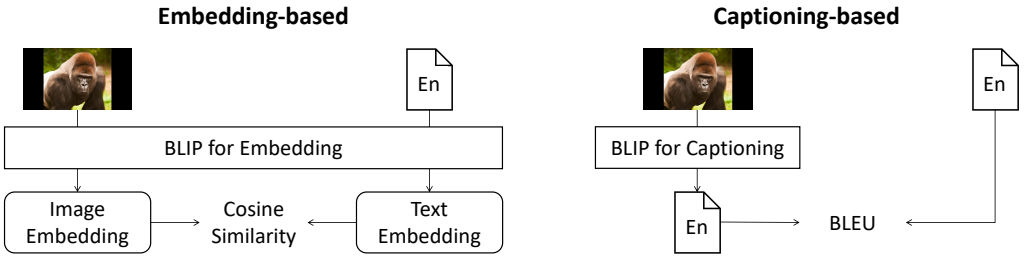


Figure 3: Overview of the proposed method of image-text matching.

start and end timestamps are recorded for each sentence (Step 2 in Figure 2), and the audio is segmented by `ffmpeg` (Step 3 in Figure 2).

Automatic Speech Recognition The audio, which was segmented into sentences in Step 3, is converted into text using `Google Speech Recognition`.[5] (Step 4 in Figure 2)

3.2 Image Annotation

In this section, we further annotate images on top of our corpus to construct five sets. Since some scenes in TED videos do not represent the content of the lecture, such as scenes showing only the speaker, we collect multiple images for each sentence. Specifically, we use three images corresponding to the beginning, middle, and end of the timestamp of each sentence. Three images per sentence were extracted using `OpenCV` library[6] with video and timestamps for each lecture video.

3.3 Parallel Corpus Filtering

The IWSLT2017 En-Ja corpus originally released 223k sentence pairs, but only 212k sentence pairs allowed us to access the videos from the URLs. To reduce noise in the corpus due to errors in timestamping and alignment, we automatically filter our parallel corpus. We filter out noisy sentence pairs by both sentence length difference and word error rate (WER) between automatically recognized (ASR) and manually transcribed (REF) English sentences. We keep $0.8 \leq \text{len(ASR)/len(REF)} \leq 1.2$ cases with small sentence length differences. Where len$(\cdot)$ is the number of words in the sentence. It also keeps WER(ASR, REF) ≤ 0.5 cases with small WER. In the WER calculation, text was lowercased and symbols were removed as a preprocessing step. This left 102k sentence pairs.

We further exclude pairs where all images are unrelated to the text, which is not beneficial to the image-guided MT. We compute the cosine similarity between the text and each of the three images assigned to it, and employ the 70,000 sentence pairs in descending order of their maximum value for our experiment. Here, BLIP-based multimodal embeddings (Li et al., 2022b) are used for similarity calculations, as in the next section.

4 Image-guided Machine Translation

In this study, as shown in Figure 1, we are given an English sentence that has been automatically recognized or manually transcribed from a lecture subtitle as well as three images that correspond in time to the text. The image-guided machine translation that we are working on is the task of inputting one image selected from among three images along with its English text and translating it into Japanese subtitles.

To select the image related to a given English sentence, we estimate the semantic similarity between vision and language. Both of the following two proposed methods are based on BLIP (Li et al., 2022b), a pre-trained multimodal model.

- **Embedding-based method:** Encode each given text and image with BLIP and then rank multiple images by the cosine similarity between their embeddings.
- **Captioning-based method:** Generate an English caption with BLIP from a given image and rank multiple images by the BLEU (Papineni et al., 2002) between the input text and the caption. (Right side of Figure 3)

5 Evaluation

5.1 Setting

Model Our multimodal MT model employed the Selective Attention model[7] (Li et al., 2022a).

[5] https://github.com/Uberi/speech_recognition
[6] https://opencv.org
[7] https://github.com/libeineu/fairseq_mmt

This model is a 4-layer, 128-dimensional Transformer (Vaswani et al., 2017) combined with image features from the Vision Transformer (vit_tiny_patch16_384) (Dosovitskiy et al., 2021). RAdam (Liu et al., 2020) was used for optimization and trained with a batch size of $4,096$ tokens and a learning rate of $1e-4$. Training was terminated when the cross-entropy loss in the validation dataset was not updated 10 times.

Data The TAIL corpus described in Section 3 was used for our experiments. We used 70,000 sentence pairs for training, 2,669 for validation, and 2,371 for evaluation. As a preprocessing, MosesTokenizer[8] (Koehn et al., 2007) and MeCab[9] (IPADIC) (Kudo et al., 2004) were used for word segmentation for English and Japanese, respectively. Subsequently, a subword segmentation with a vocabulary size of $16,000$ was performed by fastBPE[10] (Sennrich et al., 2016).

Comparison We evaluate the effectiveness of our image-text matching for image-guided MT by comparing it to the following three baseline models. Each model is trained three times with changing random seed, and the averaged BLEU (Papineni et al., 2002) is reported.

- **w/o Image baseline:** Text-only MT model. We discuss the effectiveness of the image-guided MT in comparison to this baseline.
- **w/ Random Image baseline:** An image-guided MT model that uses a randomly selected image from the entire dataset. We discuss the effectiveness of the use of related images in comparison to this baseline.
- **w/ Related Image baseline:** An image-guided MT model that uses a randomly selected image from a set of three images that correspond in time to given sentence. We discuss the effectiveness of the use of the most related images in comparison to this baseline.

5.2 Results

Automatic Evaluation Experimental results are shown in the BLEU columns of Table 1. Note that the ASR column is the translation quality for automatically recognized English sentences, while the IWSLT column is for manually transcribed English sentences. Compared to the baseline model without images, the other image-guided MT models achieved significantly higher translation quality. This suggests the effectiveness of complementing MT of lecture subtitles with images.

	BLEU		Accuracy
	ASR	IWSLT	IWSLT
w/o Image	3.94	4.73	-
w/ Random Image	7.04	8.98	-
w/ Related Image	7.07	8.97	0.495
Embedding-based	**7.30**	**9.48**	**0.785**
Captioning-based	6.96	8.90	0.410

Table 1: Performance of English-Japanese Translation.

Two baselines of image-guided MT (w/ Random Image and w/ Related Image) achieved comparable translation quality. This suggests that simply using images that correspond in time to the input text does not necessarily result in high performance. In contrast, our embedding-based method of selecting images to match text achieved the best performance for both ASR and IWSLT text.

Human Evaluation The Accuracy column in Table 1 shows the human evaluation of the accuracy of image selection for randomly sampled 200 texts. Note that these samples do not include cases where all images are related to or unrelated to the text. As with translation quality, our embedding-based method achieved the best performance. These results reveal a strong correlation between the performance of image-text matching and translation quality. It is suggested that multimodal MT performance can be improved by selecting images that are well related to the text.

6 Conclusion

In this study, we constructed a multimodal parallel corpus of images, audio in English, ASR sentences in English, transcribed sentences in English, and reference translations in Japanese of approximately 75k sentence pairs to generate cross-lingual subtitles from lecture videos. Experimental results reveal that our embedding-based image-text matching method contributes to improved performance of image-guided machine translation. Our future work includes further improvement of translation quality by combining multiple images.

[8] https://github.com/moses-smt/mosesdecoder
[9] https://taku910.github.io/mecab/
[10] https://github.com/glample/fastBPE

Acknowledgments

These research results were obtained from the commissioned research (No.22501) by National Institute of Information and Communications Technology (NICT), Japan.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1856–1862.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does Multimodality Help Human and Machine for Translation and Image Captioning? In *Proceedings of the First Conference on Machine Translation*, pages 627–633.

Iacer Calixto and Qun Liu. 2017. Incorporating Global Visual Features into Attention-based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017.

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 Ninth International Conference on Learning Representations*.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022a. On Vision Features in Multimodal Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6327–6337.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900.

Jindřich Libovický and Jindřich Helcl. 2017. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 196–202.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the Variance of the Adaptive Learning Rate and Beyond. In *Proceedings of the Eighth International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

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

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*, pages 8748–8763.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association*, pages 3655–3659.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*.

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*, 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 *Proceedings of the First Conference on Machine Translation*, pages 543–553.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal Machine Translation through Visuals and Speech. *Machine Translation*, 34:97–147.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. CoVoST: A Diverse Multilingual Speech–To-Text Translation Corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203.

Ryoya Yuasa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2023. Multimodal Neural Machine Translation Using Synthetic Images Transformed by Latent Diffusion Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 76–82.