Towards Fully Automated Manga Translation

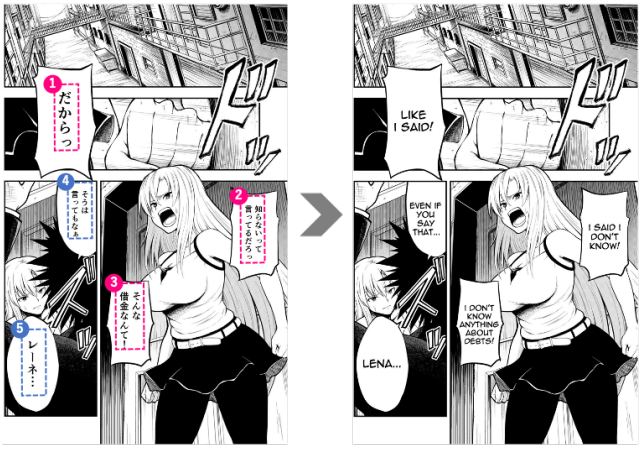
Ryota Hinami[1*], **Shonosuke Ishiwatari**[1*], **Kazuhiko Yasuda**[2], **Yusuke Matsui**[2]
[1] Mantra Inc.
[2] The University of Tokyo
{hinami,ishiwatari}@mantra.co.jp
matsui@hal.t.u-tokyo.ac.jp, yasuda@tkl.iis.u-tokyo.ac.jp

Abstract

We tackle the problem of machine translation (MT) of manga, Japanese comics. Manga translation involves two important problems in MT: context-aware and multimodal translation. Since text and images are mixed up in an unstructured fashion in manga, obtaining context from the image is essential for its translation. However, it is still an open problem how to extract context from images and integrate into MT models. In addition, corpus and benchmarks to train and evaluate such models are currently unavailable. In this paper, we make the following four contributions that establish the foundation of manga translation research. First, we propose a multimodal context-aware translation framework. We are the first to incorporate context information obtained from manga images. It enables us to translate texts in speech bubbles that cannot be translated without using context information (e.g., texts in other speech bubbles, gender of speakers, etc.). Second, for training the model, we propose the approach to automatic corpus construction from pairs of original manga and their translations, by which a large parallel corpus can be constructed without any manual labeling. Third, we created a new benchmark to evaluate manga translation. Finally, on top of our proposed methods, we devised a first comprehensive system for fully automated manga translation.

Introduction

Comics are popular all over the world. There are many different forms of comics around the world, such as manga in Japan, webtoon in Korea, and manhua in China, all of which have their own unique characteristics. However, due to the high cost of translation, most comics have not been translated and are only available in their domestic markets. What if all comics could be immediately translated into any language? Such a panacea for readers could be made possible by machine translation (MT) technology. Recent advances in neural machine translation (NMT) (Cho et al. 2014; Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015; Wu et al. 2016; Vaswani et al. 2017) have increased the number of applications of MT in a variety of fields. However, there are no successful examples of MT for comics.



Input image **Output image**

Figure 1: Given a manga page, our system automatically translates the texts on the page into English and replaces the original texts with the translated ones. ©Mitsuki Kuchitaka.

What makes the translation of comics difficult? In comics, an utterance by a character is often divided up into multiple bubbles. For example, in the manga page shown on the left side of Fig. 1, the female's utterance is divided into bubbles #1 to #3 and the male's into #4 to #5. Since both the subject "I" and the verb "know" are omitted in the bubble #3, it is essential to exploit the context from the previous or next bubbles. What makes the problem even more difficult is that the bubbles are not simply aligned from right to left, left to right, or top to bottom. While the bubble #4 is spatially closed to #1, the two utterances are not continuous. Thus, it is necessary to parse the structure of the manga to recognize the texts in the correct order. In addition, the visual semantic information must be properly captured to resolve the ambiguities. For example, some Japanese word can be translated into both "he", "him", "she", or "her" so it is crucial to capture the gender of the characters.

These problems are related to *context* and *multimodality*; considering context is essential in comics translation and we need to understand an image to capture the context. Context-aware (Jean et al. 2017; Tiedemann and Scherrer

[*]Both authors contributed equally to this work.

2017; Wang et al. 2017) and multimodal translation (Specia et al. 2016; Elliott et al. 2017; Barrault et al. 2018) are both hot topics in NMT but researched independently. Both are important in various kinds of application, such as movie subtitles or face-to-face conversations. However, there have not been any study on how to exploit context with multimodal information. In addition, there are no public corpus and benchmarks for training and evaluating models, which prevents us from starting the research of multimodal context-aware translation.

Contributions

This paper addresses the problem of translating manga, meeting the grand challenge of fully automated manga translation. We make the following four contributions that establish a foundation for research on manga translation.

Multimodal context-aware translation. Our primary contribution is a context-aware manga translation framework. This is the first approach that incorporates context information obtained from an image into manga translation. We demonstrated it significantly improves the performance of manga translation and enables us to translate texts that are hard to be translated without using context information, such as the example presented above with Fig. 1.

Automatic parallel corpus construction. Large in-domain corpora are essential to training accurate NMT models. Therefore, we propose a method to automatically build a manga parallel corpus. Since, in manga, the text and drawings are mixed up in an unstructured manner, we integrate various computer vision techniques to extract parallel sentences from images. A parallel corpus containing four million sentence pairs with context information is constructed automatically without any manual annotation.

Manga translation dataset. We created a multilingual manga dataset, which is the first benchmark of manga translation. Five categories of Japanese manga were collected and translated. This dataset is publicly available.

Fully automatic manga translation system. On the basis of the proposed methods, we built the first comprehensive system that translates manga fully automatically from image to image. We achieved this capability by integrating text recognition, machine translation, and image processing into a unified system.

Related Work

Context-Aware Machine Translation

Despite the recent rapid progress in NMT (Cho et al. 2014; Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015; Wu et al. 2016; Vaswani et al. 2017), most models are not designed to capture extra-sentential context. The sentence-level NMT models suffer from errors due to linguistic phenomena such as referential expressions (e.g., outputting 'him' when correct output is 'her') or omitted words in the source text (Voita, Sennrich, and Titov 2019b). There have been interests in modeling extra-sentential context in NMT to cope with these problems. The previously proposed methods aimed at context-aware NMT can be categorized into two types: (1) extending translation units from a single sentence to multiple sentences (Tiedemann and Scherrer 2017; Bawden et al. 2018; Scherrer, Tiedemann, and Loáiciga 2019); and (2) adding modules that capture context information to NMT models (Jean et al. 2017; Wang et al. 2017; Tu et al. 2018; Werlen et al. 2018; Voita et al. 2018; Maruf and Haffari 2018; Zhang et al. 2018; Maruf, Martins, and Haffari 2019; Xiong et al. 2019; Voita, Sennrich, and Titov 2019a,b).

While the various methods have been evaluated on different language pairs and domains, we mainly focused on Japanese-to-English translation in manga domains. Our scene-based translation is deeply related to 2+2 translation (Tiedemann and Scherrer 2017), which incorporates the preceding sentence by prepending it to be the current one. While it captures the context in the previous sentence, our scene-based model considers all the sentences in a single scene.

Multimodal Machine Translation

The manga translation task is also related to multimodal machine translation (MMT). The goal of the MMT is to train a visually grounded MT model by using sentences and images (Harnad 1990; Glenberg and Robertson 2000). More recently, the NMT paradigm has made it possible to handle discrete symbols (e.g., text) and continuous signals (e.g., images) in a single framework (Specia et al. 2016; Elliott et al. 2017; Barrault et al. 2018).

The manga translation can be considered as a new challenge in the MMT field for several reasons. First, the conventional MMT assumes a single image and its description as inputs (Elliott et al. 2016). However, manga consists of multiple images with context, and the texts are drawn in the images. Second, the commonly used pre-trained image encoders (Russakovsky et al. 2015) cannot be used to encode manga images as they are all trained on natural images. Third, no parallel corpus is available in the manga domain. We tackled these problems by developing a novel framework to extract visual/textual information from manga images and an automatic corpus construction method.

Context-Aware Manga Translation

Now let us introduce our approach to manga translation that incorporates the multimodal context. In this section, we will focus on the translation of *texts* with the help of image information, assuming that the text has already been recognized in an input image. Specifically, suppose we are given a manga page image I and N unordered texts on the page. The texts are denoted as $\mathcal{T}$, where $|\mathcal{T}| = N$. We are also given a bounding box for each text: $\mathbf{b}(t) = [x, y, w, h]^\top$. Our goal is to translate each text $t \in \mathcal{T}$ into another language t'.

The most challenging problem here is that we cannot translate each t independently of each other. As discussed in the introduction, incorporating texts in other speech bubbles is indispensable to translate each t. In addition, visual semantic information such as the gender of the character sometimes helps translation. We first introduce our approach

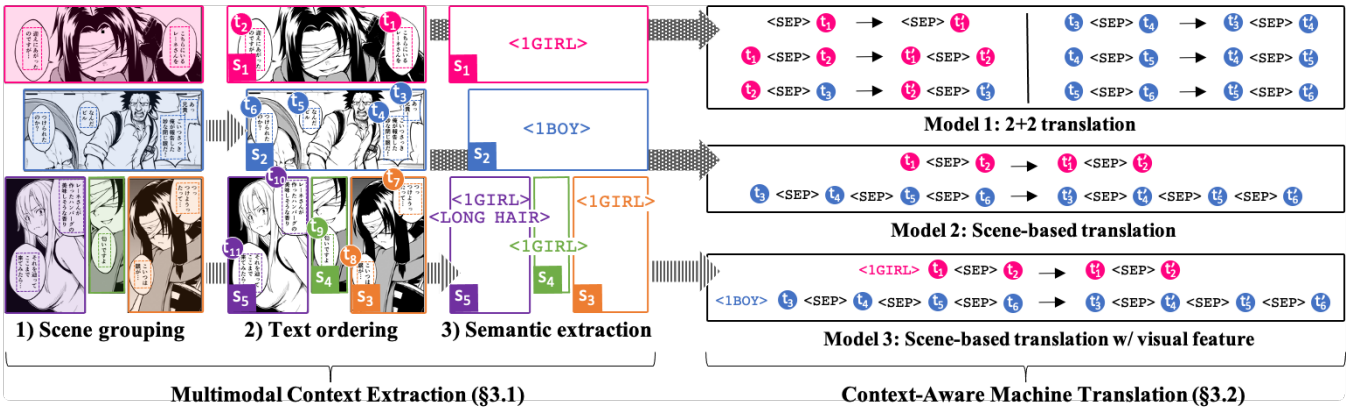


Figure 2: Proposed manga translation framework. N' represents the translation of a source sentence N. ©Mitsuki Kuchitaka

to extracting such context from an image and then describe our translation model using those contexts.

Extraction of Multimodal Context

We extract three types of context, i.e., scene, reading order, and visual information, which are all useful information for multimodal context-aware translation. The left side of Fig. 2 illustrates the three procedures explained below 1)–3).

1) Grouping texts into scenes: A single manga page includes multiple frames, each of which represents a single scene. In the translation of the story, the texts included in the same scene are usually more useful for translation than the texts in a different scene. Therefore, we group texts into scenes to determine the ones useful as contexts. First, we detect frames in a manga page using an object detector by regarding each frame as an object in the manner of (Ogawa et al. 2018). In particular, we trained the Faster R-CNN detector (Ren et al. 2015) with the Manga109 dataset (Matsui et al. 2017). Given a manga page, the detector outputs a set of scenes $\mathcal{S}$. Each scene $\mathbf{s} \in \mathcal{S}$ is represented as a bounding box of a frame: $\mathbf{s} = [x, y, w, h]^\top$. For each text $t \in \mathcal{T}$, we find the scene $\mathbf{s} \in \mathcal{S}$ that the text belongs to. Such a scene is defined as one that maximally overlaps the bounding box of the text. This is determined by an assignment function $\mathbf{a} : \mathcal{T} \to \mathcal{S}$, where $\mathbf{a}(t) = \arg\max_{\mathbf{s} \in \mathcal{S}} \mathrm{IoU}(\mathbf{b}(t), \mathbf{s})$, where IoU computes the intersection over the union for two boxes.

2) Ordering texts: Next, we estimate the reading order of the texts. More formally, we sort the unordered set $\mathcal{T}$ to make an ordered set $\{t_1, \ldots, t_N\}$ as shown in the left side of Fig. 2. Since, in manga, a single sentence is usually divided up into multiple text regions, it is quite important to ensure the text order is correct. Manga is read on a frame-by-frame basis. Therefore, the reading order of the texts is determined from the order of 1) the frames and 2) the texts in each frame. We estimate the order of the frames from the general structure of manga: *each page consists of one or more rows, each consisting of one or more columns, recursively repeating.* Each page is read sequentially from the top row, and each row is read from the right column. On the basis of this knowledge, we estimate the reading order by recursively splitting mange page vertically and horizontally. Afterward, the reading order of the texts in each frame is determined by the distance from the upper right point of each frame. Even though this approach does not use any supervised information, it accurately estimates the reading order of the frames. Some examples are shown in Fig. 3 We confirmed that it could identify the correct reading order of 91.9% of the 258 pages we tested (we evaluate with PubManga dataset introduced in the experiments section). The remaining 8.2% were irregular cases (e.g., diagonally separated, multiple frames overlapping, etc.).

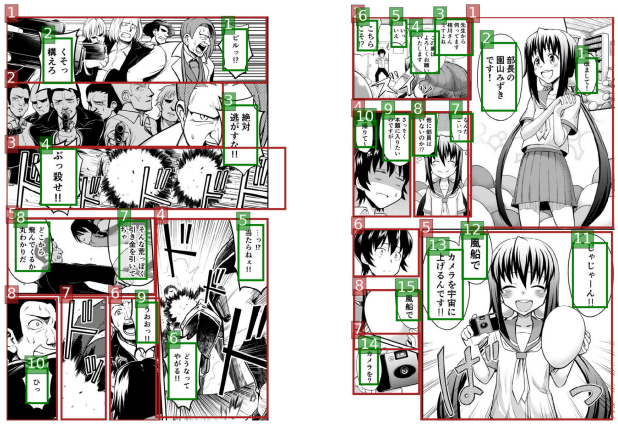


Figure 3: Results of our text and frame order estimation. The bounding boxes of text and frame are shown in the green and red rectangles, respectively. The estimated orders of text and frame are depicted at the upper left corner of bounding boxes. ©Mitsuki Kuchitaka

3) Extracting visual semantic information: Finally, we extract visual semantic information, such as the objects ap-

pearing in the scene. To exploit the visual semantic information in each scene, we predict semantic tags for each scene by using the illustration2vec model (Saito and Matsui 2015). Given a target scene $\mathbf{s} \in \mathcal{S}$, the illustration2vec module f describes the scene by predicting semantic tags: $f(\mathbf{s}) \subseteq \mathcal{L}$. In the illustration2vec model, $\mathcal{L}$ contains 512 pre-defined semantic tags: $\mathcal{L} = \{\texttt{1GIRL}, \texttt{1BOY}, \ldots\}$. Several tags can be predicted from a single scene. Although we tried integrating a deep image encoder as is done in many multimodal tasks (Zhou et al. 2018; Fukui et al. 2016; Vinyals et al. 2015), it did not improve performance on our tasks.

We should emphasize that this framework is not limited to manga. It can be extended to any kind of media having multimodal context, including movies and animations, by properly defining the scene. For example, it can be easily applied to movie subtitle translation by extracting contexts in three steps: 1) segmenting videos into scenes, 2) ordering texts by time, and 3) extracting semantic tags by video classification.

Context-Aware Translation Model

To incorporate the extracted multimodal context into MT, we take a simple yet effective concatenation approach (Tiedemann and Scherrer 2017; Junczys-Dowmunt 2019): concatenate multiple continuous texts and translate them with a sentence-level NMT model all at once. Note that any NMT architecture can be incorporated with this approach. In this study, we chose the Transformer (big) model and set its default parameters in accordance with (Vaswani et al. 2017). The right side of Fig. 2 illustrates the three models explained below.

Model1: 2+2 translation. The simplest method utilizes the previous text as context. To train and test the model, we prepend the previous text in the source and target languages (Tiedemann and Scherrer 2017). That is, to translate t_n into t'_n, two texts t_{n-1} and t_n are fed into the translation model, which outputs t'_{n-1} and t'_n. The boundary of the two texts is marked with a special token <SEP>.

Model2: Scene-based translation. Considering only the previous text as the context is not always sufficient. We may want to consider two or more previous texts or even the subsequent texts in the same scene. To enable this, we generalize the 2+2 translation by concatenating all the texts in each frame and translating them all at once. This procedure is illustrated as follows. Suppose we would like to translate t_n to t'_n. Unlike Model1 that makes use of t_{n-1} only, we feed $\{t \in \mathcal{T} \mid \mathbf{a}(t_n) = \mathbf{a}(t)\}$ into the model.

Model3: Scene-based translation with visual feature To incorporate the visual information into Model2, we prepend the predicted tags to the sequence of the input texts. Each tag is represented as a special token, such as <1GIRL> or <1BOY>. Note that this does not lead to any changes in the model itself. By adding the tags as input, we let the model consider the visual information when needed. This means that, to translate t_n into t'_n, we additionally input $f(\mathbf{a}(t_n))$.

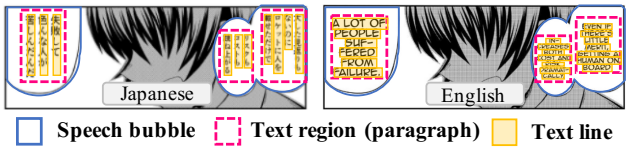


Figure 4: Term definition for manga text. ©Mitsuki Kuchitaka

Parallel Corpus Construction

We propose the approach to automatic corpus construction for training our translation model. Given a pair of manga books as input: a Japanese manga and its English-translation, our goal is to extract parallel texts with context information that can be used to train the proposed model. This is a challenging problem because manga is regarded as a sequence of images without any text data. Since texts are scattered all over the image and are written in various styles, it is difficult to accurately extract texts and group them into sentences. In addition, even when sentences are correctly extracted from manga images, it is difficult to find the correct correspondence between sentences in different languages. The differences in text direction from one language to another (e.g., vertical in Japanese and horizontal in English) makes this problem harder. We solve this problem by using computer vision techniques by fully utilizing the structural features of manga images, such as the pixel-level locations of the speech bubbles.

Terms and available labeled data First though, let us define the terms associated with manga text; Fig. 4 illustrates speech bubbles, text regions, and text lines. One speech bubble contains one or more text regions (i.e., paragraph), each comprising one or more text lines. We assume that only the annotation of speech bubbles is available for training models; annotations of text lines and text regions are unavailable. In addition, segmentation masks of speech bubbles and any data in the target language are also unavailable. This is a natural assumption because current public datasets only have speech bubble-level bounding box annotations of the Japanese version for manga (Matsui et al. 2017) and those of English version for American-style comics (Iyyer et al. 2017; Guérin et al. 2013). This limitation on labeled data is one of the challenges of parallel text extraction from comics. Note that our approach does not depend on specific languages. We also applied it to Chinese as a target language in addition to English, which is demonstrated later in Fig. 9.

Training of Detectors We train two object detectors: speech bubble and text line detectors, which is the basic building block of our corpus construction pipeline. We use Faster R-CNN model with ResNet101 backbone (He et al. 2016) for both object detectors. The object detectors are trained with the annotation of bounding boxes. While the speech bubble detector could be trained with public datasets (e.g., Manga 109), the annotations of the text lines were not available. Therefore, we devised a way to generate annotations of text lines from the speech bubble-level annota-

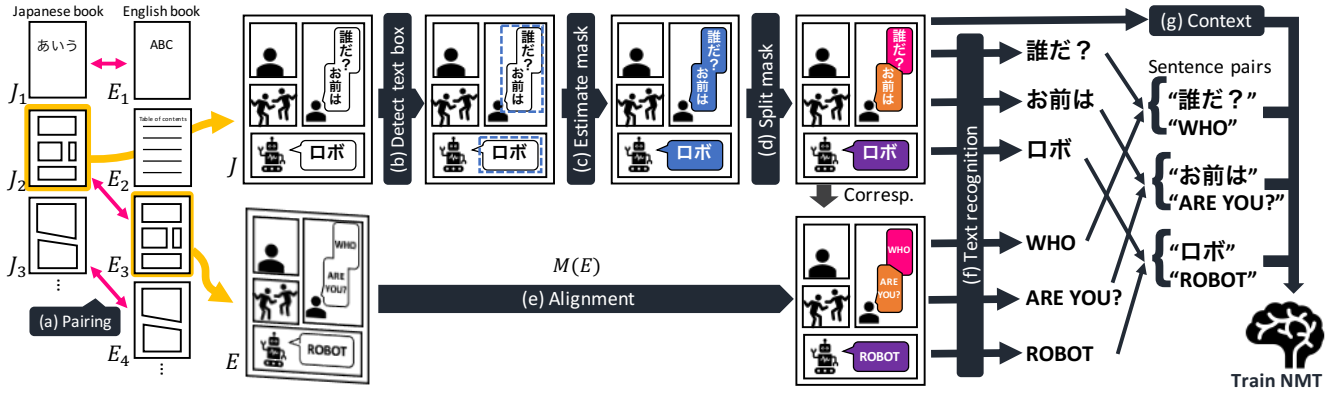


Figure 5: Proposed framework of parallel corpus construction.

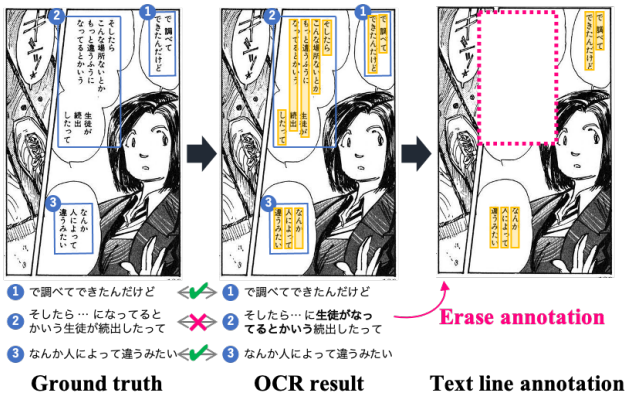


Figure 6: Generation of textline annotations. ⓒYasuyuki Ohno

tion in a weakly supervised manner. Fig. 6 illustrates the process of generating annotations. Suppose we have images with annotations of the speech bubbles' bounding boxes and texts. In this paper, we use the annotations of Manga109 dataset (Ogawa et al. 2018; Aizawa et al. 2020). We detect the text line with a rules-based approach (whose algorithm is described in supplementary material); then we recognize characters using our text line recognition module. If the recognized text perfectly matches the ground truth, we consider the text lines to be correct and use them as the annotation of text lines. Other text regions that were not recognized are filled in with white (e.g., inside the dotted rectangle in Fig. 6). The object detector is trained with the generated images and bounding box annotations. Although the rules-based approach sometimes misses complicated patterns such as bubbles containing multiple text regions, the object detector can detect them by capturing the intrinsic properties of text lines.

Extraction of Parallel Text Regions

Fig. 5 (a)–(g) illustrates the proposed pipeline for extracting parallel text regions.

(a) Pairing pages. Let us define an input Japanese manga as a set of n_j images (pages), denoted as $\{J_1, \ldots, J_{n_j}\}$. Similarly, let us define the English manga as a set of n_e pages: $\{E_1, \ldots, E_{n_e}\}$. Note that typically $n_j \neq n_e$, because pages such as the front cover, table of contents, and illustration can be optionally included or removed during the production of the translation. Owing to such inconsistencies, we must find *page-wise correspondences* first as shown in Fig. 5 (a). We find the correspondences by global descriptor-based image retrieval combined with spatial verification (Radenović et al. 2018). For each Japanese page J_i, we first retrieve the English image E_j with the highest similarity to J_i from $\{E_1, \ldots, E_{n_e}\}$, where the similarity of two pages is computed as the L_2 distance of global features extracted by the deep image retrieval (DIR) model (Gordo et al. 2016). We then apply spatial verification (Philbin et al. 2007) to reject false matching pairs. The homography matrix between two pages is estimated by RANSAC (Fischler and Bolles 1981) with AKAZE descriptors (Alcantarilla, Nuevo, and Bartoli 2013). If the number of inliers in RANSAC is more than 50, we decide that J_i and E_j are corresponding.

(b) Detection of text boxes. After the page-aligning step, we obtain a set of corresponding pairs of English and Japanese pages. Hereafter, we discuss how to extract a parallel corpus from a single pair, J and E. First, the bounding boxes of the speech bubbles are obtained by applying the speech bubble detector to J.

(c) Pixel-level estimation of speech bubbles. We estimate the precise pixel-level mask for each bubble from the bounding box. We employ edge detection with canny detector (Canny 1986) to detect the contour of speech bubbles. For each bounding box of a speech bubble, we select the connected component of non-edge pixels that shares the largest area with the bounding box, which is the blank area inside the speech bubble. In this way, we precisely estimate the masks of the speech bubbles without having to worry about how to train a semantic segmentation model that cannot be trained with the currently available dataset.

(d) Splitting connected speech bubbles. As illustrated in Fig. 4, sometimes a speech bubble includes multiple text regions (i.e., paragraphs). We split up such speech bubbles

in order to identify the text regions by clustering the text lines. The text lines obtained by the object detector are then grouped into paragraphs by clustering the vertical coordinates at the top of text lines with MeanShift (Comaniciu and Meer 2002). Finally, masks are split so that all text regions are perfectly separated, and the length of the boundary (i.e., splitting length) is minimized.

(e) Alignment between languages. We then estimate the masks of text regions for E by aligning J and E. Since the scales and margins are often different between J and E, E is transformed so that the two images overlap exactly. We update E by applying a perspective transformation: $E \leftarrow M(E)$, where $M(\cdot)$ indicates the transformation computed in the previous page pairing step. The resulting page has a better pixel-level alignment so that text regions in E can be easily localized from the text regions in J. Such a correspondence is made possible by the distinctive nature of manga: the translated text is located in the same bubble. Note that we do not use any learning-based models for E in steps 1)–5), so our method can be used for any target language even if a dataset for learning detectors is unavailable.

(f) Text recognition. Given the segmentation masks of the text regions, we recognize the characters for each image pair J and E. Since we found that existing OCR systems perform very poorly on manga text due to the variety of fonts and styles that are unique to manga text, we developed an OCR module optimized for manga. We developed our own text rendering engine that generates text images optimized for manga. Five millions of text images are generated with the engine, by which we train the OCR module based on the model of Baek et al. (Baek et al. 2019). Technical details of this component are described in the supplementary material.

(g) Context extraction. We extract the context information (i.e., the reading order and scene labels of each text) from J in the manner described in the previous section.

Experiments

Dataset

Although there are no manga/comics datasets comprising of multiple languages, we created two new manga datasets, i.e., **OpenMantra** and **PubManga**, one to evaluate the MT, the other to evaluate the constructed corpus.

OpenMantra: While we need a ground-truth dataset to evaluate the NMT models, no parallel corpus in the manga domain is available. Thus, we started by building OpenMantra, an evaluation dataset for manga translation. We selected five Japanese manga series across different genres, including fantasy, romance, battle, mystery, and slice of life. In total, the dataset consists of 1593 sentences, 848 frames, and 214 pages. After that, we asked professional translators to translate the whole series into English and Chinese. This dataset is publicly available for research purposes.[1]

PubManga: OpenMantra is not appropriate for evaluating the constructed corpus because translated versions are created by ourselves. Thus, we selected nine Japanese manga series across different categories, each having 18–40 pages (258 pages in total), and created another dataset of published translations (PubManga). This dataset includes annotations of 1) bounding boxes of the text and frame, 2) texts (character sequence) in both Japanese and English, and 3) the reading order of the frames and texts. The annotations and full list of manga titles are available upon request.

Evaluation of Machine Translations

To confirm the effectiveness of our models and Manga corpus, we ran translation experiments on the OpenMantra dataset.

Training corpus: To train the NMT model for manga, we collected training data by the proposed corpus construction approach. We prepared 842,097 pairs of manga pages that were published in both Japanese and English. Note that all the pages are in digital format without textual information. 3,979,205 pairs of Japanese–English sentences were obtained automatically. We randomly excluded 2,000 pairs for validation purposes.

In addition, we used OpenSubtitles2018 (OS18) (Lison, Tiedemann, and Kouylekov 2018), a large-scale parallel corpus to train a baseline model. Most of the data in OS18 are conversational sentences extracted from movie and TV subtitles, so they are relatively similar to the text in manga. We excluded 3K sentences for the validation and 5K for the test and used the remaining 2M sentences for training.

Methods: Table 1 shows the six systems used in our evaluation. **Google Translate**[3] is an NMT system used in several domains, but the sizes and domains of its training corpus have not been disclosed. We chose the **Sentence-NMT (OS18)** as another baseline. The model is trained with the OS18 corpus; therefore, there are no manga domain texts included in its training data. The **Sentence-NMT (Manga)** was trained on our automatically constructed Manga corpus described in the previous section. **Sentence-NMT (OS18)** and **Sentence-NMT (Manga)** use the same sentence-level NMT model.

While the first three systems are sentence-level NMTs, the fourth to sixth ones are proposed context-aware NMT models. We set **2 + 2** (Tiedemann and Scherrer 2017) (Model1) as the baseline and compared their performance with those of our **Scene-NMT** models with and without visual features (Model3 & Model2, respectively).

Evaluation procedure: Manga translation differs from plain text translation because the content of the images influences the "feeling" of the text. To examine how readers actually feel when reading a translated page, we conducted a manual evaluation of translated pages instead of plain texts. We recruited En–Ja bilingual manga readers. They were given a Japanese page and translated English ones, and they were asked to evaluate the quality of the translation of each

[1] https://github.com/mantra-inc/open-mantra-dataset

[3] https://translate.google.com/

System	Training corpus	Translation unit	Human	BLEU
Without context				
Google Translate[2]	N/A	sentence	-	8.72
Sentence-NMT (OS18)	OpenSubtitles2018	sentence	2.11	9.34
Sentence-NMT (Manga)	Manga Corpus	sentence	2.76	14.11
With context				
2 + 2 (Tiedemann and Scherrer 2017)	Manga Corpus	2 sentences	2.85	12.73
Scene-NMT	Manga Corpus	frame	**2.98***	12.65
Scene-NMT w/ visual	Manga Corpus	frame	**2.91***	12.22

Table 1: System description and translation performances on the OpenMantra Ja–En dataset. * indicates the result is significantly better than Sentence-NMT (Manga) at $p < 0.05$.

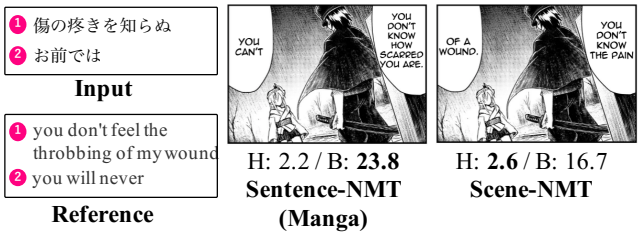


Figure 7: Outputs of the sentence-based (center) and frame-based (right) models. The values after **H** and **B** are respectively the human evaluation and BLEU scores for each page. ©Mitsuki Kuchitaka

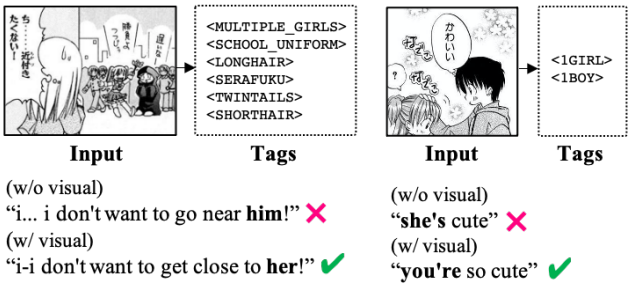


Figure 8: Translation results with and without visual features. ©Miki Ueda, ©Satoshi Arai.

English page. Following the procedure in the Workshop on Asian Translation (Nakazawa et al. 2018), we asked five participants to score the texts from 1 (worst; less than 20% of the important information is correctly translated) to 5 (best; all important information is correctly translated). All the methods explained above other than **Google Translate** were compared. The order of presenting the methods was randomized. In total, we collected 5 participants $\times$ 5 methods $\times$ 214 pages $= 5,350$ samples. See the supplemental material for the details of the evaluation system. We also conducted an automatic evaluation using the BLEU (Papineni et al. 2002).

Results: Table 1 shows the results of the manual and automatic evaluation. The huge improvement of the **Sentence-NMT (Manga)** over **Google Translate** and **Sentence-NMT (OS18)** indicates the effectiveness of our strategy of Manga corpus construction.

A pair-wise bootstrap resampling test (Koehn 2004) on the results of the human evaluation shows that the **Scene-NMT** outperformed the **Sentence-NMT (Manga)**. On the other hand, there is no statistically significant difference between **2 + 2** and **Sentence-NMT (Manga)**. These results suggest that not only the contextual information but also the appropriate way to group them is essential for accurate translation.

In contrast to the results of the human evaluation, the BLEU scores of the context-aware models (fourth to sixth lines in Table 1) are worse than that of **Sentence-NMT (Manga)**. These results suggest that the BLEU is not suitable for evaluating manga translations. Fig. 7 shows an example where the **Scene-NMT** outperformed **Sentence-NMT (Manga)** in the manual evaluation but had lower BLEU scores. Here, we can see that only the **Scene-NMT** has swapped the order of the texts. This flexibility naturally resolves the differences in word order between Japanese and English. However, it results in a worse BLEU score since the references usually maintain the original order of the texts.

Although there is no statistically significant difference between **Scene-NMT** and **Scene-NMT w/ visual**, Fig. 8 shows some promising results; pronouns ("you" and "her") that cannot be estimated from textual information are correctly translated by using visual information. These examples indicate that we need to combine textual and visual information to appropriately translate the content of manga. However, we found that a large portion of the errors of **Scene-NMT w/ visual** are caused by the incorrect visual features. To fully understand the impact of the visual feature (i.e., semantic tags) on translation, we conducted an analysis in Fig. 10: (i) and (ii) in the figure show the outputs of the **Scene-NMT** and **Scene-NMT w/ visual**, respectively. The pronoun errors in (ii) are caused by the incorrect visual feature "Multiple Girls" extracted from the original image. When we overwrote the character face with a male face, **Scene-NMT w/ visual** output the correct pronouns, as shown in (iii). This result proved that **Scene-NMT w/ visual** model consider visual information to determine translation results, and it would be improved if we devise a way to extract visual fea-

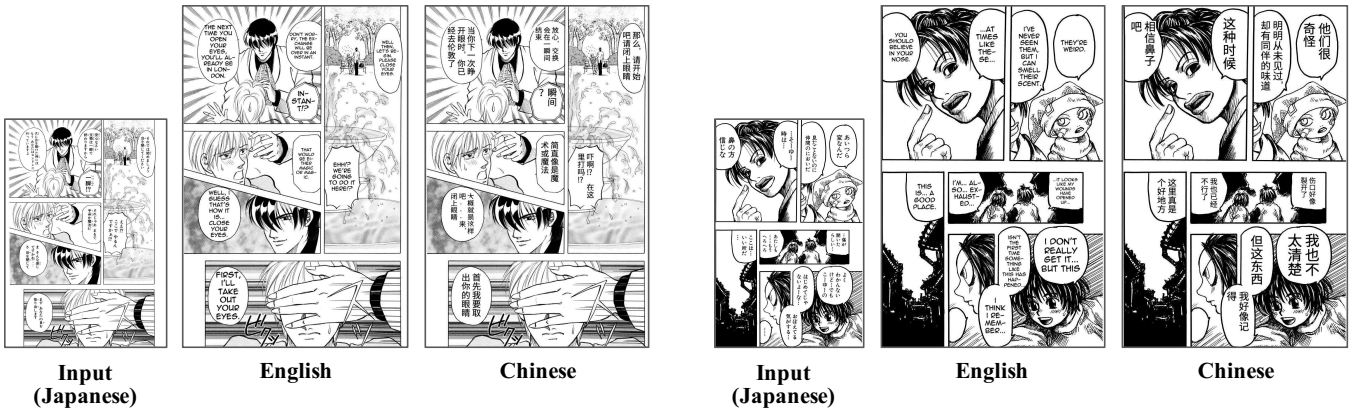


Figure 9: Results of fully automatic manga translation from Japanese to English and Chinese. ©Masami Taira, ©Syuji Takeya

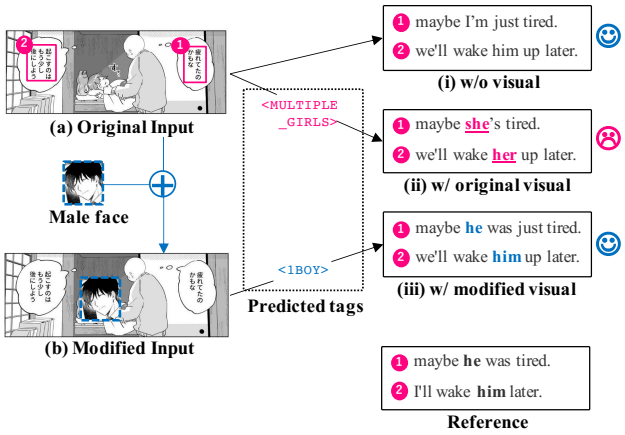


Figure 10: Translations output by the sentence-based model without (i) and with visual information (ii). By overwriting the character face in the input image (a) with a male face (b), the pronouns in the translation results (iii) are also changed. ©Nako Nameko

tures more accurately. Designing such a good recognition model for manga images remains as future work.

Evaluation of Corpus Construction

To evaluate the performance of corpus construction, we compared the following four approaches: 1) **Box**: Bounding boxes by the speech bubble detector are used as text regions instead of segmentation masks. This is the baseline of a simple combination of speech bubble detection and OCR. 2) **Box-parallel**: Bounding box of speech bubbles are detected in both Japanese and English images by applying detector to both images. For each detected Japanese box, the English box that overlaps it most is selected as the corresponding box. 3) **Mask w/o split**: Segmentation masks of speech bubbles are estimated, but the process of splitting the masks is not done. 4) **Mask w/ split** (the full proposed method): Segmentation masks of speech bubbles are estimated, and connected bubbles are split. This fully utilizes

	$> 90\%$ match		$> 70\%$ match	
Method	Recall	Prec.	Recall	Prec.
Box	0.267	0.365	0.434	0.594
Box-parallel	0.246	0.614	0.289	0.722
Mask w/o split	0.381	0.522	0.480	0.657
Mask w/ split	**0.584**	**0.653**	**0.688**	**0.769**

Table 2: Corpus construction performance on the PubManga.

the structural feature of the manga images. In 1) and 2) the regions of bounding boxes are regarded as the mask of text regions.

The corpus construction performances were evaluated on the PubManga dataset; the results are listed in Tab. 2. For a $> 90\%$ and $> 70\%$ match, the text pair with a normalized edit distance (Karatzas et al. 2013) between the ground truth and extracted texts of more than 0.9 and 0.7 were considered true positives, respectively; this allowed for some OCR mistakes because the accuracy of the OCR module is not the main focus of this experiment. This result shows that our approach that uses mask estimation is significantly better than the two approaches that use only bounding-box regions. Mask splitting also significantly improved both precision and recall. The bounding box-based approaches fail to identify the regions of English text, especially when the shapes of the text regions are different from those of the Japanese text; this problem is caused by the difference in text direction. These results indicate that parallel corpus extraction from manga cannot be done with the simple combination of OCR and object detection; exploiting structural information manga is effective. Note that we use the same OCR and detection modules in these experiments. The details of the evaluations are provided in the supplementary material.

[3]We ran the test on Apr. 17, 2020.

Fully Automated Manga Translation System

We launch a fully automated manga translation system on top of the proposed model trained with the constructed corpus. Given a Japanese manga page, the system automatically recognizes texts, translates them into target language, and replaces the original texts with the corresponding translated texts. It performs the following steps.

1) Text detection and recognition: Given a Japanese input page, the system recognizes texts in the same way as in the corpus construction. This step predicts masks of the text regions and Japanese texts with their contexts.

2) Translation: Japanese texts are translated into the target languages by using the trained NMT model. Since our approach to translation and corpus construction does not depend on a specific language, we can translate the Japanese text into any target language if unlabeled manga book pairs for constructing corpus are available.

3) Cleaning: The original Japanese texts are removed from the translation. We employ an image inpainting model for this; the regions of text lines are replaced by the inpainting model, by which texts are removed clearly even when they are on image texture or drawing. We used edge-connect (Nazeri et al. 2019), because its edge-first approach is very good at complementing defects of drawings.

4) Lettering: Finally, the translated texts are rendered with optimized font size and location on the cleaned image. The location is one that maximizes the font size under the condition that all texts are inside the text region.

Examples. Fig. 9 shows the translations produced by our system. It demonstrates that our system can automatically translate Japanese manga into English and Chinese.

Conclusion & Future Work

We established a foundation for the research into manga translation by 1) proposing multimodal context-aware translation method and 2) automatic parallel corpus construction, 3) building benchmarks, and 4) developing a fully automated translation system. Future work will look into 1) an image encoding method that can extract continuous visual information that helps translation, 2) an extension of scene-based NMT to capture longer contexts in other scenes and pages, and 3) a framework to train the image recognition models and the NMT model jointly for more accurate end-to-end performance.

Acknowledgements

This work was partially supported by IPA Mitou Advanced Project, FoundX Founders Program, and the UTokyo IPC 1st Round Program.

The authors would like to appreciate Ito Kira, Mitsuki Kuchitaka, and Nako Nameko for providing their manga for research use, and Morisawa Inc. for providing their font data. We also thank Naoki Yoshinaga and his research group for the fruitful discussions before the submission. Finally, we thank the anonymous reviewers for their careful reading of our paper and insightful comments.

References

Aizawa, K.; Fujimoto, A.; Otsubo, A.; Ogawa, T.; Matsui, Y.; Tsubota, K.; and Ikuta, H. 2020. Building a Manga Dataset" Manga109" with Annotations for Multimedia Applications. *IEEE MultiMedia* .

Alcantarilla, P. R.; Nuevo, J.; and Bartoli, A. 2013. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In *Proc. BMVC*.

Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S. J.; and Lee, H. 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proc. ICCV*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. ICLR*.

Barrault, L.; Bougares, F.; Specia, L.; Lala, C.; Elliott, D.; and Frank, S. 2018. Findings of the Third Shared Task on Multimodal Machine Translation. In *Proc. WMT*.

Bawden, R.; Sennrich, R.; Birch, A.; and Haddow, B. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proc. NAACL*.

Canny, J. 1986. A computational approach to edge detection. *IEEE TPAMI* (6): 679–698.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proc. EMNLP*.

Comaniciu, D.; and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE TPAMI* 24(5): 603–619.

Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; and Specia, L. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proc. WMT*.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proc. Workshop on Vision and Language*.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24(6): 381–395.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proc. EMNLP*.

Glenberg, A. M.; and Robertson, D. A. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language* 43(3): 379–401.

Gordo, A.; Almazán, J.; Revaud, J.; and Larlus, D. 2016. Deep image retrieval: Learning global representations for image search. In *Proc. ECCV*.

Guérin, C.; Rigaud, C.; Mercier, A.; Ammar-Boudjelal, F.; Betet, K.; Bouju, A.; Burie, J.-C.; Louis, G.; Ogier, J.-M.; and Revel, A. 2013. eBDtheque: A Representative Database of Comics. In *Proc. ICDAR*.

Harnad, S. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1-3): 335–346.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. CVPR*.

Iyyer, M.; Manjunatha, V.; Guha, A.; Vyas, Y.; Boyd-Graber, J.; III, H. D.; and Davis, L. S. 2017. The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives. In *Proc. CVPR*.

Jean, S.; Lauly, S.; Firat, O.; and Cho, K. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135* .

Junczys-Dowmunt, M. 2019. Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation. In *Proc. WMT*.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *Proc. ICDAR*.

Koehn, P. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. EMNLP*.

Lison, P.; Tiedemann, J.; and Kouylekov, M. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In *Pro. LREC*.

Maruf, S.; and Haffari, G. 2018. Document Context Neural Machine Translation with Memory Networks. In *Proc. ACL*.

Maruf, S.; Martins, A. F.; and Haffari, G. 2019. Selective Attention for Context-aware Neural Machine Translation. In *Proc. NAACL*.

Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; and Aizawa, K. 2017. Sketch-based Manga Retrieval using Manga109 Dataset. *Multimedia Tools and Applications* 76(20): 21811–21838.

Nakazawa, T.; Sudoh, K.; Higashiyama, S.; Ding, C.; Dabre, R.; Mino, H.; Goto, I.; Pa Pa, W.; Kunchukuttan, A.; and Kurohashi, S. 2018. Overview of the 5th Workshop on Asian Translation (WAT2018). In *Proc. WAT*.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Ogawa, T.; Otsubo, A.; Narita, R.; Matsui, Y.; Yamasaki, T.; and Aizawa, K. 2018. Object Detection for Comics using Manga109 Annotations. *CoRR* abs/1803.08670.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*.

Radenović, F.; Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proc. CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proc. NIPS*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 1–42.

Saito, M.; and Matsui, Y. 2015. Illustration2vec: a semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*, 1–4.

Scherrer, Y.; Tiedemann, J.; and Loáiciga, S. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proc. DiscoMT*.

Specia, L.; Frank, S.; Sima'an, K.; and Elliott, D. 2016. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proc. WMT*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence Learning with Neural Networks. In *Proc. NIPS*.

Tiedemann, J.; and Scherrer, Y. 2017. Neural Machine Translation with Extended Context. In *Proc. DiscoMT*.

Tu, Z.; Liu, Y.; Shi, S.; and Zhang, T. 2018. Learning to remember translation history with a continuous cache. *TACL* 6: 407–420.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Proc. NIPS*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proc. CVPR*.

Voita, E.; Sennrich, R.; and Titov, I. 2019a. Context-Aware Monolingual Repair for Neural Machine Translation. In *Proc. EMNLP-IJCNLP*.

Voita, E.; Sennrich, R.; and Titov, I. 2019b. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proc. ACL*.

Voita, E.; Serdyukov, P.; Sennrich, R.; and Titov, I. 2018. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proc. ACL*.

Wang, L.; Tu, Z.; Way, A.; and Liu, Q. 2017. Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proc. EMNLP*.

Werlen, L. M.; Ram, D.; Pappas, N.; and Henderson, J. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proc. EMNLP*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144.

Xiong, H.; He, Z.; Wu, H.; and Wang, H. 2019. Modeling coherence for discourse neural machine translation. In *Proc. AAAI*.

Zhang, J.; Luan, H.; Sun, M.; Zhai, F.; Xu, J.; Zhang, M.; and Liu, Y. 2018. Improving the Transformer Translation Model with Document-Level Context. In *Proc. EMNLP*.

Zhou, M.; Cheng, R.; Lee, Y. J.; and Yu, Z. 2018. A Visual Attention Grounding Neural Model for Multimodal Machine Translation. In *Proc. EMNLP*.