Large Language Models as Manga Translators: A Case Study

Zhishen Yang[1], Tosho Hirasawa[2], Edison Marrese-Taylor[3, 4], Naoaki Okazaki[1]
School of Computing, Tokyo Institute of Technology[1]
Tokyo Metropolitan University[2], The University of Tokyo[3]
National Institute of Advanced Industrial Science and Technology[4]
zhishen.yang@nlp.c.titech.ac.jp, hirasawa-tosho@ed.tmu.ac.jp
edison.marrese@aist.go.jp, okazaki@c.titech.ac.jp

Abstract

Originating in Japan, manga has gained immense popularity on a global scale as a distinct form of comics. However, the primary language of manga, Japanese, is a barrier to its widespread access to international markets. While crucial for its global expansion, manga translation is often accompanied by substantial investment in time and resources. The emergence of machine translation technology presents the opportunity to automate manga translation processes, potentially reducing translation costs. However, challenges arise due to copyright restrictions limiting training data access. This paper empirically explores the feasibility of leveraging Large Language Models (LLMs) for manga translation tasks. Our study delves into investigations to assess to what extent LLMs are capable of performing such a task and identify which contextual cues help enhance the quality of the output. The experimental results demonstrate the potential of employing the LLM in manga translation, indicating a promising trajectory for future research.

1 Introduction

Comics are one of the most famous artistic expressions worldwide. Among the diversity of comics around the world, Japanese manga emerges as one of the most prolific variations with global impacts. In this context, language barriers have represented a substantial challenge in allowing manga to reach international markets, as manual translation is usually highly time-consuming and costly. This means that many comics have not been translated and are only available in their domestic markets. Automating manga translations to other languages will greatly facilitate the accessibility and popularity of Japanese manga, ultimately leading to reduced publishing timelines and shorter wait times for readers. Although recent progress in Machine Translation (MT) has landed a promising foundation for manga translation, many challenges still need to be addressed despite this advantage.

Compared with traditional text-only MT tasks, the manga translation endeavor introduces unique challenges, including the following.

1. Utilizing visual scenes in translations As a type of visual storytelling, texts are combined with manga images to compose a complete story. Visual content often presents backgrounds, emotions, body language, and texts to immerse readers. While human translators possess the innate ability to incorporate semantic and emotional messages from visual content when translating texts, enabling MT models to effectively integrate visual scenes to improving translation quality remains a challenging and under-explored task.

2. Lack of training data Manga books are subject to copyright restrictions, and annotating manga for building translation models is highly costly and labor-intensive, leading to training data for this task being limited. This contrasts with recent trends in machine learning, with models requiring substantial amounts of data for training. In this sense, we think that the absence of readily available and high-quality manga translation training data may potentially hinder the development of successful machine translation models.

3. Language style The language of manga is artistic and narrates stories to its readers. Translation models trained on data from different domains could potentially misinterpret this creative language, misrepresenting the intended emotions and causing the inadvertent loss of the authors' implicit messaging.

Given these difficulties, tackling the task of manga machine translation continues remains an ongoing challenge. Despite the recent advances in our field with the advent of Large Language Models (LLMs), which have demonstrated a wide variety of capabilities across several natural language tasks, their effectiveness in more concrete applications, such as ours, remains unexplored. In light of this issue, we conduct empirical investigations to study the performance of LLMs in automatic manga translation. To the best of our knowledge, this study is the first to focus on employing LLMs for manga translation. Our main findings are as follows.

- Our experimental results demonstrate that the strong zero-shot translation abilities of LLM are, to a large extent, preserved when performing manga translation, demonstrating the potential of using such models for this task.
- We found that giving few-shot examples as context to the LLM improves translation quality, which lies in accordance with observations for several other tasks.
- We show that our zero-shot LLM-based approach can attain comparable performance to model fine-tuning using millions of examples, suggesting a potential direction to overcome data availability problems.

2 Proposed Approach

This study tries to answer two main questions: (1) Can LLM perform manga translation without access to the training dataset? and (2) What types of context bring performance boosts to LLM?

Given the aforementioned research questions, we first assess the effectiveness of Large Language Models in zero-shot manga translation, contrasting it with a supervised machine translation model. Following this, we investigate the importance of integrating various types of contextual information to support the LLM in enhancing translation quality. In this context, we defined three categories of context:

1. **Few-shot context:** Instances of randomly-sampled translation examples.
2. **Local context:** Adjacent textual content to the utterance to be translated.
3. **Visual context:** Textual descriptions of manga images/frames accompanying the utterance to be translated, which we first obtain using a pre-trained vision-and-language model.

3 Experimental Settings

This section details our models, dataset, and evaluation frameworks used in the experiments. Please refer to the appendix for implementation details and prompt templates.

Dataset The OpenMantra dataset [1] is the only manga translation dataset available with public access. It contains five Japanese manga books with 1,592 text passages and 214 pages in five different genres. Professional translators translated Japanese texts into English and Chinese.

Translation Direction The source language in our experiments is Japanese, and the target languages are Chinese or English.

Translation models We utilized large multilingual MT models trained on massive parallel corpora, specifically, M2M100 [2]. Additionally, we experimented with M2M100 models of different sizes, including the 418M, 1.2B and 12B models. We also experiment with LLM-based black-box chat agents, specifically GPT-3.5 [3, 4, 5].

Evaluation We performed evaluations in terms of the quality of the generated translations. Specifically, we followed the methodology of previous studies, employing BLEU-4 scores [6]. Additionally, we present results for METEOR [7], COMET [8], and BERTScore [9].

4 Results

Our study seeks to empirically investigate the viability of utilizing LLMs for manga translation, with a specific focus on identifying the contexts that contribute to enhancing translation quality.

The experimental results indicate that GPT-3.5 has strong zero-shot multilingual translation capabilities, which surpass the M2M100 models. Furthermore, few-shot contexts help the GPT-3.5 further boost its performance.

Drawing from our experimental findings, it becomes apparent that while manga translation is a type of multimodal machine translation, incorporating visual contexts has a minimal, if not slightly detrimental, impact on performance. In the next subsections, we give details of our experimental results and provide in-depth analysis to answer our research questions.

4.1 Supervised Machine Translation Model: M2M100

Table 1 presents the results of automatic evaluation metrics of M2M100, a multilingual machine translation model. We evaluated three different sizes of the model: 418M (small), 1.2B (medium), and 12B (large). Notably, the medium size variant M2M100 (1.2B) achieved the highest scores across BLEU, METEOR, chrF, and COMET metrics for both Chinese and English translations.

Based on the experimental results, scaling the model does not continually improve the scores. Using a large M2M100 model (12B) slightly decreases scores for the English translation, while we can find moderate deterioration for the Chinese translation. M2M100 (1.2B) also ranks the best according to the COMET score for English and Chinese translations. For the BERTScore, M2M100 (1.2B) performs the best on Chinese translations, although the larger model (M2M100 12B) has the best BERTScore for English translation, it is only a marginal improvement compared to M2M100 (1.2B).

The experimental results indicate that scaling does not consistently yield positive outcomes for M2M100 in the context of manga translation tasks. This is probably caused by the fact that manga texts are outside the data domains in which M2M100 was trained. Regarding target languages for translation, M2M100 consistently performs better when English is the designated target language.

4.2 Large Language Model: GPT3.5

As a Large Language Model with robust zero-shot capabilities across various natural language tasks, GPT-3.5's potential to address the manga translation task was the primary focus of this study. As shown in Table 1, even in the absence of contextual cues, GPT-3.5 offer significantly superior scores across all metrics in comparison to M2M100. Unlike the latter, GPT-3.5 achieved higher scores in translating Chinese, according to the BLEU metric, although similar improvements were not observable in other metrics.

These findings underscore GPT-3.5's potential as a large language model to undertake the manga translation task in a zero-shot scenario. We refer to GPT-3.5 without any contexts as the baseline for the rest of the experiments.

4.3 Probing the need for textual contexts

Having identified the kinds of LLM models that can obtain better performance in our task, we now task ourselves with answering our second research question, namely, what types of context bring performance boosts to the LLM (GPT-3.5). We first investigate the necessity of incorporating the local context. In the case of English translation, it is evident that including local context leads to improvements across all metrics. However, we observe a slight decline in scores when introducing a local context for the Chinese translation.

We also observe that randomly sampled utterance-level translation demonstrations (k-shot contexts) enhances the performance of the LLM. In the context of manga translation, our results show that presenting examples of manga translations potentially imparts GPT-3.5 insights into manga translation styles. To gain insight into how the model can leverage this context for better performance, we experiment by feeding randomly selected $k = [1, 2, 4, 6, 8, 10]$ utterances in Japanese along with their corresponding translations from four other manga books, thus avoiding leakage problems. We hypothesize that offering GPT-3.5 more examples should lead to improved performance.

Referencing Table 2, introducing more demonstrations does not necessarily correlate with performance improvements for both languages. In the case of English, k=10 yielded the highest BLEU, METEOR, chrF scores, while k=6 achieved the best BERTScore. On the contrary, for Chinese, when k=4, GPT-3.5 achieved the highest scores in BLEU, METEOR.

4.4 Probing the need for visual contexts

Finally, we seek to answer whether visual context contributes to translations. To that end, we used the LLaVA model to obtain textual descriptions of manga images. Given that each manga image consists of frames, we also obtained textual descriptions for each frame. This approach can be likened to an alternative to image descriptions, resembling a more focused rendition of the image context. The underlying hypothesis is that the model would provide more intricate descriptions, effectively delivering a detailed form of visual context and ultimately improving translation quality.

Table 1: Automatic evaluation results on Japanese to English/Chinese Translation, where * indicates statistical significance of the difference over GPT-3.5 ($p \leq 0.05$) bootstrap resampling, Rules indicate that we provide additional instructions on how to perform translations, Demos indicates our best results using k-shot (k=6 for English, k=4 for Chinese) contexts, $\text{Img}_{En,Zn}$ denotes models that receive visual context at the utterance level, and $\text{Frame}_{En,Zn}$ denotes models that received visual context at the frame level.

Model (JA-EN/ZH)	BLEU	METEOR	chrF	COMET	BERTScore (F1)
Supervised					
M2M100 (418M)	5.47/ 4.24	23.39/ 16.28	21.63/ 12.27	-0.47/ -0.06	**88.62**/86.15
M2M100 (1.2B)	**6.35/ 5.78**	**24.21/ 19.73**	**23.16/ 13.88**	**-0.40/ 0.05**	87.77/ **87.78**
M2M100 (12B)	6.29/ 4.70	23.06/ 17.71	22.22/ 13.17	-0.44/ -0.01	88.00/86.71
Zero-shot/Few-shot					
GPT3.5-Turbo	9.77/ 11.29	35.40/ 32.63	32.15/ 21.82	-0.04/ 0.36	89.80/ 90.27
- Rules	8.05*/ 8.46*	34.34/ 30.46*	31.16/ 20.23*	-0.10/ 0.24	89.39/ 89.74
+ Local Context	10.23/ 10.74	36.45*/ 32.50	32.97*/ 21.87	-0.02/ 0.35	89.78/ 90.19
+ Demos.	**10.76*/ 12.26**	**36.54/ 33.35**	**32.99/ 22.59**	**0.00/ 0.36**	**90.06/ 90.34**
+ Img_{En}	9.86/ 10.27*	35.84/ 31.93*	32.31/ 21.36	-0.05/ 0.32	89.04/ 89.83
+ Frame_{En}	9.99/ 10.37	35.58/ 32.40	32.20/ 21.69	-0.05/ 0.34	88.66/ 89.79
+ Img_{Zh}	9.48/ 9.70*	35.02/ 31.12*	31.51*/ 20.79*	-0.07/ 0.29	89.10/ 89.39
+ Frame_{Zh}	9.73/ 10.20	35.32/ 31.18*	31.85/ 20.90	-0.06/ 0.31	89.02/ 89.60

Table 2: Ablation study: role of k in few-shot experiments in Japanese to English/Chinese Translations, where k denotes the number of shots.

k	BLEU	METEOR	chrF	COMET	BERTScore
1	10.04/ 11.47	35.21/ 32.31	31.97/ 22.05	-0.03/ **0.37**	89.47/ 90.24
2	10.72/ 11.52	35.93/ 32.35	32.73/ 21.76	-0.01/ 0.35	89.79/ 89.98
4	10.43/ **12.26**	36.24/ **33.35**	32.64/ 22.59	**0.00**/ 0.36	90.03/ **90.34**
6	**10.76**/ 11.23	36.54/ 32.81	32.99/ 22.22	**0.00**/ 0.35	**90.06**/ 90.22
8	10.47/ 12.07	36.25/ 33.2	32.71/ **22.63**	-0.01/ 0.36	89.92/ 90.31
10	**10.76**/ 11.87	**36.57**/ 33.21	**33**/ 22.47	**0.00**/ 0.36	90.00/ 90.30

Considering that target translations are in English and Chinese, we hypothesize that providing visual contexts written in the target languages would offer more informative cues for the model to utilize in the translation process.

From the experimental results in Table 1, for translation into English, English image description helps slightly improve BLEU, METEOR, and chrF scores but not significantly. Still, we cannot observe the same improvements when translating into Chinese, which results in drops in all metrics. Compared with Chinese image descriptions, English image descriptions always gain higher evaluation metric scores.

Overall, we see that using frame description does not improve the evaluation metric scores. GPT-3.5 obtained similar scores for both translation directions compared to image descriptions. Surprisingly, using image and frame descriptions in English always leads to higher evaluation scores for both English and Chinese translations.

Compared to our baseline GPT-3.5 model, integrating frame descriptions did not improve the evaluation metric scores for both translation directions. In both cases, GPT-3.5 obtained scores similar to those obtained using image descriptions. However, for both translation directions, using frame descriptions led to slight improvements in BLEU.

5 Conclusion

In this paper, we carried out empirical investigations into applying LLMs for the task of manga translation. In this pursuit, we selected GPT3.5 as our LLM. We observed that LLMs exhibit robust zero-shot capabilities in translating manga texts. Furthermore, our findings indicate that introducing few-shot context in the form of translation examples, enhances the translation quality of GPT-3.5.

This study marks an initial exploration into utilizing LLMs for the manga translation task. The complexity inherent in manga translation encompasses multiple layers, including modelling and resource building. In future research, we intend to discover the feasibility of fine-tuning LLMs or vision-language models to address the manga translation challenge. Additionally, we are intrigued by the investigation of which visual elements within manga images contribute to improving translation quality.

Acknowledgment

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

References

[1] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards fully automated manga translation. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 12998–13008, 2021.

[2] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. Beyond English-Centric Multilingual Machine Translation. **Journal of Machine Learning Research**, Vol. 22, No. 107, pp. 1–48, 2021.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[4] Introducing ChatGPT.

[5] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 3008–3021. Curran Associates, Inc., 2020.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[7] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[8] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[9] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **International Conference on Learning Representations**, September 2019.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, April 2023.

A Implementation Details

In experiments, we utilized the M2M100 models, which are accessible through HuggingFace and the Microsoft Azure GPT services [1)]. Since the ground truth Chinese translations are in traditional characters, we converted them to simplified characters. For the the vision and language model used in the experiments, we employed LLaVa [2)] [10] specifically the version Vicuna-13B-v1.3-CLIP-L-336px, to generate textual descriptions for images/frames. The English prompt used was "Please describe this manga image", while the corresponding Chinese prompt was "请用中文描述 这个漫画 图片".

B Prompt template

```
You are a translation assistant from {src_lang} to {tgt_lang}.
Some rules to remember:
- Maintaining the contents' accuracy is important, but since texts are from manga, we want
 ↪ to prioritize naturalness and ease of communication.
- Instead of translating word by word, try to translate the whole sentence or phrase at
 ↪ once.
- Number of translated sentences should be the same as the number of input sentences.
- Return translations without additional explanations, comments, or source inputs in
 ↪ {src_lang}.
{context_type}:
{context}
Input format:
1. Sentence
Please translate the text followed by the above format into {tgt_lang}:
{text}
Please return translations in the following format:
1. Translation
```

Figure 1: The prompt template used in experiments.

Figure 1 shows the prompt template for GPT3.5. Where {text} represent nth sentence to be translated: {src_lang} represents source language: Japanese, {tgt_lang} denotes target language: Chinese or English. {context_type} is a sentence that explicitly tells GPT3.5 the context type, as shown below:

1. **Few-shot context:** Here are sample translations:
2. **Local context:** Empty
3. **Visual context:** Here is the manga image description:

{context} denote context content

1. **Few-shot context:** K translation examples.
2. **Local context:**
   Previous sentence:
   n-1th sentence
   Next sentence:
   n+1th sentence
3. **Visual context:** Image/frame descriptions

1) https://azure.microsoft.com/en-us/products/ai-services/openai-service
2) https://github.com/haotian-liu/LLaVA