The Efficacy of Pre-editing as a Means of Improving NMT Output

Kayo Tsuji, Osaka Metropolitan University, Japan
Benjamin Neil Smith, Kwansei Gakuin University, Japan
Kiyo Okamoto, Osaka Metropolitan University, Japan

The Asian Conference on Education 2024
Official Conference Proceedings

Abstract
Neural machine translation (NMT/MT) replicates human brain networks and is often considered a "black box" due to its unexplained structure and function. To use NMT effectively, target-language (TL) learners should pre-edit source-language (SL) documents, similar to how humans paraphrase first-language (L1) texts before translating to their second language (L2) (Tsuji, 2024). Previous studies have shown that pre-editing SL texts improves the quality of MT output. However, the specific differences between NMT output with and without editing have been explored less. This study aims to compare NMT translations with and without pre-editing and examine the impact of pre-editing on the quality of TL texts. The study involved 54 Japanese university students with intermediate English proficiency, each composing a Japanese text which was translated into English using NMT. Three language researchers analysed issues in the unedited NMT output, and systematically categorised them into lexical, grammatical, semantic and formatting errors. Comparing the results with those of Tsuji (2024), exploring issues in pre-edited NMT outputs, showed that NMT output *without* pre-editing displayed more frequent errors than NMT output *with* pre-editing. In particular, the semantic issues in unedited NMT output significantly limited readability. On the contrary, the pre-edited NMT translations described in Tsuji (2024) were generally comprehensible and contained minor errors that did not significantly affect comprehensibility. This study indicates that pre-editing SL texts is essential for improving the quality of NMT translations. The errors illustrated in this study could potentially be used to assist students learning L1 paraphrasing skills required for the human L2 translation process.

Keywords: Pre-editing, Neural Machine Translation, NMT Errors, Japanese-English Translation

iafor
The International Academic Forum
www.iafor.org

Introduction

Neural machine translation (NMT) has significantly advanced in recent years, achieving higher accuracy compared to statistical machine translation (SMT) (Bahdanau et al., 2015; Sutskever et al., 2014). Designed to mimic human neural networks, NMT continues to evolve, enhancing its capabilities (Ninomiya et al., 2021). Despite its progress, the internal workings of NMT systems remain largely opaque, often referred to as a "black box." Nonetheless, translations produced by modern NMTs are widely perceived to approximate human translation quality.

To further enhance NMT quality, human intervention in the form of pre-editing (modifying the source text) and post-editing (refining the output text) is often required. With the focus of this study being pre-editing, this paper begins with an exploration of four studies that highlight its effectiveness (Farhana et al., 2023; Feifei et al., 2022; Kokanova et al., 2022; Liang & Han, 2022). Following this is an examination of two studies reporting some of its potential limitations (Marzouk & Hansen-Schirra, 2019; Rantan, 2024) with a final discussion of research addressing practical considerations for pre-editing. While prior studies offer mixed insights into the efficacy of pre-editing in improving MT quality, the continuous evolution of NMT raises the question of whether pre-editing will remain necessary as the technology advances.

Previous Studies

Value of Pre-editing

Several studies have demonstrated the advantages of pre-editing for enhancing NMT output. For example, Farhana et al. (2023) investigated English translations of Indonesian texts, focusing on students in translation and interpreting courses as participants. Their findings revealed that over 90% of the pre-edited NMT output was error-free, producing clear and comprehensible English translations. Key pre-editing strategies included reordering sentence elements, refining vocabulary, and eliminating superfluous phrases, leading to improved accuracy and efficiency. The authors emphasized the importance of pre-editing training to maximize its benefits.

Similarly, Liang and Han (2022) compared the different benefits of pre-editing and post-editing when translating academic and medical texts between English and Chinese. They found pre-editing to be essential for enhancing semantic accuracy, while post-editing helped to refine linguistic and cultural aspects of the output text that MT lacks awareness of. These findings underscore the complementary roles of pre- and post-editing in producing high-quality translations.

Feifei et al. (2022) examined the impact of pre-editing on translating an English academic book into Chinese. Their study revealed significant improvements in BLEU scores due to precise adjustments such as adding, omitting, or replacing words and ensuring technical terms were accurately rendered, resulting in contextually appropriate translations.

In another study, Kokanova et al. (2022) evaluated the translation of English news articles from *bbc.com* into Russian. They observed that pre-editing reduced lexical inaccuracies, particularly in handling polysemous words, which were often mistranslated without pre-

editing. This intervention enhanced both clarity and precision, highlighting the value of pre-editing in mitigating common translation errors.

Collectively, these studies demonstrate the substantial role of pre-editing in improving MT quality, particularly in addressing lexical challenges and refining text clarity across various genres.

The Potential Limitations of Pre-editing

That being said, not all research supports the efficacy of pre-editing. Rantan (2024) analysed English-to-Finnish translations, specifically focusing on the challenges posed by Finnish prepositions. The findings indicated that pre-editing strategies were insufficient to fully address errors associated with complex sentence structures and, consequently, the study suggested that post-editing might be more effective for such cases.

Marzouk and Hansen-Schirra (2019) evaluated the application of controlled language (CL) rules in German-to-English translations of technical manuals. Their results showed minimal differences in error rates before and after applying CL rules, suggesting that modern NMT systems can often produce accurate texts without extensive pre-editing. They concluded that the relevance of pre-editing might diminish as NMT technology advances, and this would likewise diminish the need to investigate and implement effective strategies thereof.

Despite the benefits of pre-editing, these studies raise questions about its necessity in an era of increasingly sophisticated NMT systems.

Purpose of the Study

This study aims to identify differences in translation quality between pre-edited and unedited source texts (STs), focusing on Japanese-to-English translations of academic reports written by Japanese learners of English. Specifically, it seeks to verify the efficacy of pre-editing as a means to improve NMT output. The central research question is: What are the specific differences between NMT output produced with and without pre-editing? By addressing this question, the study could illuminate the areas where pre-editing enhances translation accuracy and readability in this context, as well as its potential limitations.

Methodology

This study compares the NMT output without pre-editing to the output with pre-editing. To achieve this, data from Tsuji (2024) was used as pre-edited NMT output. In said study, 73 students with intermediate English proficiency wrote reports in Japanese during the spring semester of 2022. The participants input their texts without pre-editing into a NMT system to generate NMT output. They then compared the resulting target-language texts (TTs) to the source-language texts (STs), identifying parts of the translations that did not accurately reflect the ST. The problematic parts in STs were subsequently rewritten (pre-edited), and then re-entered into NMT. The students repeated this until they were satisfied that the ST meaning was accurately reflected in their TTs. The final TTs, consisting of about 15-20 sentences, and the revised STs were collected as raw data. Each student selected several TSs that they were unable to verify the accuracy of, making for a total of 183 TL sentences.

Reflecting the above, the current study involved 54 students who wrote academic reports in Japanese during the spring semester of 2023. They followed the same procedure and identified parts of the translations that did not accurately reflect the STs. However, this set did not pre-edit their sentences and submitted them as is. Each TT consisted of approximately 400 words, organized into 15-20 sentences, with 144 TL sentences being submitted for analysis in total. In identifying the common errors that students couldn't self-diagnose, typical MT errors present in students' writing can be illustrated. In both cases, the STs and TTs were collected as raw data.

Participants were all students enrolled in English courses taught by the researchers, and provided consent for their writing to be used for research purposes. The participants were deemed appropriate as they had an intermediate level of English proficiency, representative of a typical university student in Japan. As in Tsuji (2024), the 54 participants were instructed to create their academic texts on their own specialised subject under the condition that their degree of familiarity with the topic should be relatively high.

Three language researchers, comprising one native English speaker and two native Japanese speakers, analysed both sets of the data. The analysis was conducted manually by the analysts initially working independently and later collaborating to compile and classify results based on mutual agreement. The analysts identified syntactic and semantic errors, as well as mistranslations, in the unedited MT output, and categorized these problematic elements accordingly.

Among the freely available online NMT tools, this study focused on DeepL, which has demonstrated strong performance based on BLEU (Bilingual Evaluation Understudy) scores, a widely recognized metric for assessing MT quality (Fujii et al., 2021). Higher BLEU scores indicate translations that are closer in quality to human-generated translations.

Results

As in Tsuji (2024), three language researchers systematically categorised errors in unedited NMT output into *Lexical, Grammatical, Semantic* and *Formatting Issues*. *Lexical Issues* included inappropriate use of TL vocabulary and redundant or repeated TL expressions, while the *Grammatical Issues* discovered were mainly those such as misuse or lack of determiners, misuse or lack of prepositions, and inappropriate singular and plural forms. *Semantic Issues* related to incomprehensible sentences and missing information. Finally, *Formatting Issues* indicated misuse of punctuation, misuse of upper- and lower-case characters, and missing quotation marks.

The findings of this investigation reveal that many errors were present in the unedited translations, while pre-edited NMT translations displayed fewer and relatively minor errors, shown in Tsuji (2024).

Two prominent error categories, in particular, relating to *Lexical* and *Semantic Issues*, showed significant difference between the NMT output with and without pre-editing.

With regard to *Lexical Issues*, there was little difference between the pre-edited and unedited NMT output. Moreover, these errors largely did not impede the comprehensibility of the sentences in both the NMT output with and without pre-editing. This issue may therefore

require greater TL (English) knowledge to repair, as pre-editing had little effect on its occurrence.

Regarding *Semantic Issues*, without pre-editing, errors emerged more frequently, negatively impacting the accuracy and readability of the translations. In contrast, pre-editing let authors simplify sentence structures, thereby reducing semantic ambiguities.

The errors in the unedited NMT output were frequently linked to long or overly complex SL sentences. TL sentences observed in the unedited output mixed multiple pieces of information in a single sentence, and the subject and predicate did not always align, making it difficult to understand the relationship between each piece of information or what was and was not a necessary sentence element. Further analysis highlighted the nature of these semantic issues, which were especially prevalent in the unedited NMT translations.

The results emphasize the critical role of pre-editing in enhancing the clarity and precision of NMT output.

Examples of Lexical Issues

Detailed examples of *Lexical Issues* are discussed in this section.

NMT Output Without Pre-editing

Example 1
MT output[1]: There is a problem that children with *disabilities float* in class due to their *disability* characteristics.

The word “float” does not make sense in this context, wherein the author meant that children with disabilities may be left behind or ignored in the class. Moreover, “disability” is incorrectly used as an adjective, where the correct phrasing ought to be “disability’s characteristics” or, potentially, “disabled characteristics”. Since “disabilities” is mentioned earlier in the sentence, the word could also be deleted and the sentence would make more sense. That said, despite these problems the overall meaning of the sentence comes through in the NMT output.

Example 2
MT output[2]: Although I was still not very good at *making* hairpins and smashes, when I was able to actually *make* them in the matches, I *could feel* the results of my technical practice, and I *felt* a sense of accomplishment.

Although there is nothing mistaken in this sentence, it displays repetition of the verbs “make” and “feel” in different forms and its length makes its meaning somewhat difficult to parse. By using synonyms or equivalent phrases or by separating the ideas into different sentences, its readability would be improved.

[1] The SL equivalent of the MT output is as follows: 障害特性により、障害を持った子供たちがクラスで浮いてしまうという問題がある。
[2] ヘアピンやスマッシュの精度はまだまだでしたが、実際に試合で決めることができたときは、技術練習の成果を感じることができ、達成感を感じました。

NMT Output With Pre-editing

Example 3
MT output[3]: A semiconductor made from a single crystal with no impurities is called a *true* semiconductor, in which the outermost electrons of the atoms that make up the semiconductor, *such as silicon,* are used for covalent bonding without excess or deficiency.

Here, the term "true semiconductor" is incorrectly used in place of the term "intrinsic semiconductor", and the placement of the example "such as silicon" would work better following the initial term it is an example of. Despite this, the overall meaning does come through, albeit somewhat awkwardly.

Example 4
MT output[4]: In addition, companies are set up to conduct development *locally*. However, their arrangement for R&D activities is still less than that of companies that do manufacturing *locally* or sales *locally*.

These sentences use the word "locally" repetitively and although the meaning is clear its readability would be improved by either cutting one instance of the word (for example, "manufacturing or sales locally") or using a synonym in its place.

From these examples, it can be seen that both the NMT output with and without pre-editing exhibited lexical issues, but this did not tend to affect their comprehensibility in both cases. Within this data, this type of issue can be edited during the phase of post-editing, but given its minor nature it is less of a priority for NMT users to focus on.

Examples of Semantic Issues

Detailed examples of *Semantic Issues* are discussed in this section.

NMT Output Without Pre-editing

Example 5
MT output[5]: It can be inferred that the element sought in American *anime* is the ability to *experience* through images what one cannot *experience* oneself by looking down from a third person's point of view, to feel *a sense of* elation and to *experience a sense of* the extraordinary.

The organization of the text in a relatively long sentence make the author's meaning unclear, as it is difficult to understand exactly how one point connects to another. Moreover, repetitive phrasing ("experience", "a sense of") and the wrong term make it more unintelligible. The use of "anime" is incorrect in this context, as this term is used mainly for animations made in

[3] 不純物を持たない単結晶で作られた半導体を真性半導体といい、シリコンなどの半導体を構成する原子の最外殻電子が過不足なく共有結合に使用された状態である。
[4] また、開発を現地で行う会社が設置される。しかし、研究開発活動を行うためのそれらの配置は、製造を現地でする会社や販売を現地でする会社に比べるとまだ少ない。
[5] 第三者の視点から俯瞰して見ることで、自分では体験できないことを、映像を通して体験することで高揚感を抱いたり、非日常感を味わったりすることができるのがアメリカのアニメに求められる要素なのではないかと推測できる。

Japan with a particular and iconic style. A more appropriate term would be “animations” or “cartoons”.

Example 6
MT output[6]: I believe that *women, among all the stressful household chores*, lack understanding and appreciation from those around them.

This example illustrates incorrect word order and semantic ambiguity. The phrase “women, among all the stressful household chores” suggests that women are somehow included as a household chore, which is nonsensical. This error stems from ambiguity in the ST, which could have been resolved by reconstructing the sentence during the pre-editing phase.

NMT Output With Pre-editing

Example 7
MT output[7]: The background of the incident may be that people in a socially distressed position, of low status and low income, were not in a situation where they could fully use their *reason* due to the hardships of life and dissatisfaction with the *disparity*.

In this sentence, there is missing information, since the form of “disparity” is unmentioned (for example, it may refer to “wealth disparity”). Moreover, the NMT has incorrectly used the word “reason” rather than “reasoning”. However, given the proximity of “reason” to “reasoning” and the context within the sentence, the true meaning can be easily gleaned without much effort from the reader.

Example 8
MT output[8]: Bischofia javanica is a popular guide item *as a “bleeding tree”*; when the bark is injured, red sap flows from the wound.

This sentence is also missing information, as the subject is “known as a bleeding tree” (rather than literally being a “bleeding tree”) due to its characteristics. As with example 7, however, this too, can be understood by the reader due to the context provided in the sentence.

From these examples, it is clear that pre-editing can help reduce the impact of semantic problems and improve the overall comprehensibility of sentences. The output of unedited STs was often long with ideas mixed together in ways that made the overall meaning difficult to parse, whilst the output from pre-edited STs – although still displaying errors – could be understood without difficulty.

Examples of Other Frequent Errors in Unedited NMT Output

Here are sample sentences of frequent errors: inappropriate use of subjects, inappropriate use of conjunctions, and illogical causal relationships.

---

[6] 私は女性が、家事をストレスに感じる中でも特に周りからの理解や感謝が足りないのではないかと思う。
[7] 事件の背景にあるのは、社会的に苦しい立場にいる、低い地位、低収入の人々が、生活苦や格差に対する不満などで、理性を十分に働かせるような状況になかったということでないかと思われる。
[8] アカギとは「血を流す木」として人気のあるガイド項目で、樹皮を傷つけると傷口から赤い樹液が流れる。

Inappropriate Use of Subjects

Other frequently observed errors fell under the category of inappropriate use of subjects. Japanese learners of English often omit the subject in their writing, as it is typically implied in Japanese and contextually understood (Tsuji, 2024). This characteristic led to frequent subject-related errors.

Example 9
MT output[9]: Therefore, *we* decided to prevent influenza infection by increasing the rate of hand disinfection using bottles of disinfectant ethanol, which *we* are tempted to push.

In this example, the double use of "we" is inappropriate as it creates ambiguity – either the people who decided to use disinfectant ethanol are the same as those who are tempted to use it, or these are two distinctly defined groups of people. Given the context of preventing the spread of infection, it is likely that the second "we" is referring to a larger group (which may or may not include those of the former "we"). Therefore, substituting this for a generic term such as "people" would improve intelligibility and make it clear that the latter "we" is separate and greater, and that the former "we" were making a decision for the benefit of a larger group. This more closely aligns it with the ST.

Based on the findings of Tsuji (2024), pre-editing had a significant effect on this kind of issue. As mentioned, Japanese is a language that often omits explicit subjects in favor of implied ones, which can lead to ambiguous or inappropriate translations when entered into NMT systems. Pre-editing addressed this by allowing the author to explicitly supplement implied subjects, improving clarity and syntactic correctness.

Inappropriate Use of Conjunctions

Other frequent errors, pertaining to inappropriate use of conjunctions, were also prominent, as illustrated by the following sentence.

Example 10
MT output[10]: I have not had much opportunity to work with people who lack self-care at this point, *and* I do not think I have been able to develop such skills.

Although grammatically and structurally correct, the conjunction "and" does not effectively convey the causal relationship implied in this sentence. Replacing it with "so" would better reflect the intended connection between the two statements. Notably, the ST itself lacked punctuation, which may have contributed to the suboptimal translation. This could easily be remedied through pre-editing, giving the author an opportunity to add the necessary punctuation to help clarify meaning.

Illogical Causal Relationships

Another notable issue was illogical causal relationships.

---

[9] そこで、つい押したくなる消毒用エタノールのボトルを使用し、手指消毒の実施率をあげることでインフルエンザ感染を予防することとした。
[10] 私自身もあまり現時点でセルフケアが欠如している人と関わる機会がなくそういった力をつけることができていないと思うので。

Example 11
MT output[11]: This has increased the income of farmers, and the population has also risen due to the *increase in society*.

This output demonstrates a logical inconsistency. The statement “the population has also risen due to the increase in society” is vague and illogical, as the rising population would likely cause societal changes rather than result from them. While the translation is grammatically accurate, the logical disconnect likely originates from the ST. Such issues must be addressed during pre-editing to ensure coherent causal relationships in the NMT output.

Conclusion

This study sought to identify specific differences in NMT output with and without pre-editing. Analysis revealed that errors relating to lexical and semantic problems were particularly prevalent in translations without pre-editing. Semantic errors, such as incomprehensible sentences and missing information, were observed frequently and had significant implications for the comprehensibility of a sentence. NMT output without pre-editing often exhibited poorly expressed sentence elements, illogical idea connections, and subject inference errors, resulting in unintelligible texts. However, the degree and frequency of these errors in pre-edited NMT outputs are reduced to the point that they did not significantly hamper readability, demonstrating its effectiveness in enhancing clarity and accuracy of NMT output. Whatever developments improve NMT capabilities, the results suggest that, at present, pre-editing remains essential to improve the quality of NMT output.

Limitations

Despite its valuable insights, the study had several limitations. The sample consisted of a limited number of students with intermediate TL proficiency. This was due to the small-scale of the study, as the participants were all students enrolled in English courses taught by the researchers. Moreover, the pre-edited NMT output data was not collected since the time allocated to complete the task was limited by the curriculum prescribed by the university administration. Furthermore, the study focused exclusively on the NMT output of academic reports translated from Japanese as a source language to English as a target language. These limitations mean that the findings may not generalize to other language pairs, genres of writing, different levels of TL proficiency in the authors, or alternative contexts. Future research should involve larger participant groups, adopt a longitudinal design, and include multiple genres to provide more comprehensive insights.

[11] それにより農家の収入も上がり、社会増加による人口の上昇も実現した。

References

Bahdanau, D., Cho, KH., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations, USA*, 1–15. https://doi.org/10.48550/arXiv.1409.0473

Farhana, B. C. D., Baharuddin, W. A. L., & Farmasari, S. (2023). Academic text quality improvement by English department students of University of Mataram: A study on pre-editing of Google neural machine translation. *Jurnal Ilmiah Profesi Pendidikan*, *8*(1), 247–254. http://doi.org/10.29303/jipp.v8i1.1186

Feifei, F., Rong, C., & Xiao, W. (2022). A study of pre-editing methods at the lexical level in the process of machine translation. *Arab World English Journal for Translation & Literary Studies*, *6*(2), 54–69. https://doi.org/10.31235/osf.io/3yrej

Fujii, R., Mita, M., Abe, K., Hanawa, K., Morishita, M., Suzuki, J., & Inui, K. (2021). Phenomenon-wise evaluation dataset towards analyzing robustness of machine translation models. *Natural Language Processing*, *28*(2), 450–478. https://doi.org/10.5715/jnlp.28.450

Kokanova, E. S., Berendyaev, M. V., & Kulikov, N. Y. (2022). Pre-editing English news texts for machine translation into Russian. *Language Studies and Modern Humanities*, *4*(1), 25–30. https://doi.org/10.33910/2686-830X-2022-4-1-25-30

Liang, Y., & Han, W. (2022). Source text pre-editing versus target text post-editing in using Google Translate to provide health services to culturally and linguistically diverse clients. *Science, Engineering and Health Studies*, *16*, 1–5. https://doi.org/10.14456/sehs.2022.25

Marzouk, S., & Hansen-Schirra, S. (2019). Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation*, *33*, 179–203. https://doi.org/10.1007/s10590-019-09233-w

Ninomiya, T., Deguchi, Y., Uchiyama, M., Tamura, A., Sumida, A. (2021). Studies on bilingual subword segmentation for neural machine translation. *Japio Year Book 2021*, 276–285. https://japio.or.jp/00yearbook/files/2021book/21_4_03.pdf

Rantan, M. (2024). *Neural machine translation and Finnish case-inflections: Translation and problems and pre-editing possibilities.* [Master thesis, University of Turku]. Turun Yliopisto. https://www.utupub.fi/handle/10024/178423

Sutskever, I., Vinyals, O., & Le, V. Q. (2014). Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems, Canada*, *2*, 3104–3112. https://doi.org/10.48550/arXiv.1409.3215

Tsuji, K. (2024). Identifying MT errors for higher-quality target language writing. *International Journal of Translation, Interpretation, and Applied Linguistics*, *6*(1), 1–17. http://doi.org/10.4018/IJTIAL.335899