# Pre-editing Plus Neural Machine Translation for Subtitling: Effective Pre-editing Rules for Subtitling of TED Talks

Yusuke Hiraoka
Kansai University
yusuke@wakate-honyaku.net

Masaru Yamada
Kansai University
yamada@apple-eye.com

## Abstract

In this study, the authors developed a set of pre-editing rules for TED Talk subtitling to translate Japanese source text into English. The simplified rules optimized for NMT (@TexTra® Minnano jido hon'yaku) were intended for use by a monolingual pre-editor of content to be disseminated in English. The rules were a) insert punctuation b) make implied subjects and objects explicit, and c) write proper nouns in English. The effectiveness of the rules was evaluated by human raters and BLEU score. Quality improvement was confirmed significant on human evaluation, although in some cases no changes or even degrade in quality were observed. However, one of the main concerns about the feasibility of this approach, the 21-character limit specified in the TED subtitling guidelines, was validated. The authors hold that pre-editing plus NMT is a promising approach to translating TED Talk subtitles.

## 1 Introduction

The translation quality of neural machine translation (hereafter referred to as NMT or MT) has improved drastically when compared with that of previous systems such as statistical machine translation. NMT systems have been in practical use for the English-Japanese combination since 2016. This technological advancement is expected to help ease the effects of a worldwide shortage of translators. To fully meet the increasing translation demand of all language combinations would require approximately two billion translators (Common Sense Advisory, 2018). NMT, with its advantages in cost and delivery time, could be a solution for this excessive demand.

While post-editing has already established itself as a means of translation for specific purposes in industry, pre-editing has not yet been in practice. Research on pre-editing is also under development, especially for the English-Japanese combination. Pre-editing cost-effectiveness and effective pre-editing strategies have not yet been investigated (Miyata & Fujita, 2017).

Despite the lack of evidence or precedent supporting the adoption of pre-editing, the authors of the present study see ample potential in it, particularly monolingual pre-editing, by which a person with limited knowledge of the target language (i.e. English) would be able to publish their own content translated from their L1 (Japanese).

The development of information technology has also impacted the translation process. For instance, the source content of audiovisual translation has become multifaceted, ranging from user-generated videos such as YouTube[1] to TED Talks[2]. This shift has led to increased demand for subtitling with low cost and quick turnaround. Ultimately, it would be ideal if content creators or non-professional fansub translators could perform pre-editing of their own content for dissemination.

Given this background, the present study will investigate the possibility of monolingual pre-editing of online audiovisual contents, specifically TED talk subtitling, by non-

[1] https://www.youtube.com/
[2] https://www.ted.com/

professional translators—from Japanese into English—with an aim to establish a set of effective monolingual pre-editing rules which is considered to be easily adopted by online communities.

## 2 Research question

The study aims to develop and test a set of simple, effective pre-editing rules for audio-visual contents including TED Talk subtitling to translate Japanese source text to English, using @TexTra® Minnano jido hon'yaku,[3] an NNT engine developed by the National Institute of Information and Communications Technology (NICT) in Japan. The three rules are based on the previous research and are intended to be as simple and easy to follow as possible, so they can be used by monolingual users with limited knowledge of English. Therefore, the present study will examine the following question: how effective are those pre-editing rules at improving NMT quality of TED Talk subtitling?

## 3 Experiment design

### 3.1 TED subtitling

TED Talks is a free online video service led by TED, a non-profit organization, that promotes a global TED conference where a number of well-known speakers deliver presentations on "ideas worth spreading," normally in the English language. Under the umbrella of TED, the organization also holds regional conferences worldwide at which local speakers present in their native language.

The source content to be investigated in this study was a presentation from TEDxTokyo 2012 to 2013, delivered in Japanese and transcribed by volunteer TED Talk viewers. These transcriptions became the source texts fed into an NMT system to be translated into English. For analysis, this set of source texts, pre-edited by the researcher and machine-translated, was used to compare the final quality.

### 3.2 Text type of the source speech

The TED source speech content for this study is a presentation delivered in Japanese, transcribed in the original language, and then translated into multiple languages by volunteer translators in the TED translation project.

The Japanese texts used for the present study were transcriptions of excerpts from four videos shown in Appendix B. The entire text data comprise approximately 12,000 Japanese characters with 606 subtitle segments in total.

In accordance with TED subtitling guidelines, these transcribed subtitles contain sound representations (e.g. "laugh" and "applause") for enhanced accessibility to deaf and hard-of-hearing viewers which are not normally seen in professionally-produced movie subtitles. Thus, for the present investigation, these were omitted prior to the comparative analysis.

### 3.3 TED subtitles as target text

The English target text of the TED presentation to be used as a reference point for quality evaluation was translated by TED volunteer translators. It contains approximately 4,900 English words with 616 subtitle segments in total. The translation quality of volunteer-created subtitles is regarded to be close to the professional quality because TED volunteer translators have to go through a rigid translation process involving multiple reviews, and they are required to follow TED-specific subtitling guidelines, including the following rules:

1. keep the subtitle reading speed at a maximum of 21 characters per second (CPS);
2. try to preserve as much meaning as possible.

These rules are different from conventional movie subtitling norms that limit characters to under 12 CPS, which is approximately half the number of characters allowed in TED subtitles. The looser character limit adopted in the TED subtitling may relate to viewers being able to rewind the video and watch portions they missed again. The liberalized character limit also allows TED subtitling to make more 'literal' translations than in conventional movie subtitling so it can better preserve the source meaning. Conventional movie subtitling with the 12-character limit normally requires editing and condensing source information to fit through 'sense-based' translation or trans-creation. Therefore, these

[3] https://mt-auto-minhon-mlt.ucri.jgn-x.jp

TED subtitling rules—permitting more characters and more literal translation that aims to preserve the source meaning—are considered favorable for the use of MT, and worth investigating.

### 3.4 Pre-editing method

Pre-editing is generally categorized into two methods, bilingual pre-editing and monolingual pre-editing. Bilingual pre-editing allows the pre-editor to edit the source text while looking at the MT output whereas monolingual pre-editing does not. Thus, monolingual pre-editing requires no target language skill.

The focus of this research is monolingual pre-editing since part of our ultimate goal is to enable content creators or people with limited target language command (i.e. monolingual speakers) to pre-edit the source text of their own language to disseminate content. For this purpose, it is desirable to set simple pre-editing rules for pre-editors to follow.

### 3.5 Monolingual pre-editing rules

Hiraoka & Yamada (2019) previously carried out an investigation to create pre-editing rules for popular Japanese YouTube content and selected the top 19 most effective editing categories in terms of quality improvement.

From the 19 pre-editing rules, the authors of this study chose three to observe (Table 1) based on frequency (cf. Miyata & Fujita, 2017) and ease of use, considering the potential post-editor to be a non-bilingual content creator with limited knowledge of the target language and also low editing skills in the source language.

As Miyata & Fujita (2017) states, pre-editing normally requires skillful editing of the source language to identify and edit errors that violate rules provided in the specific instructions. Thus, for this investigation we have selected a very simple set of rules that monolingual speakers can follow easily without referring to the target language.

| Rule | Type | Method |
|---|---|---|
| **1** | Punctuation | Compensate missing punctuation (tôten) |
| **2** | Subject / Object | Compensate missing subject and/or object |
| **3** | Proper Noun | Write proper nouns in target language (English) |

Table1. Pre-edit Rules

As shown in Appendix C, rules include 1) inserting missing punctuations based on spaces, line breaks and segment breaks of the original source texts, 2) compensating subjects and/or objects of the sentence since the Japanese language is a pro-drop language in which certain pronouns are omitted when they are pragmatically or grammatically inferable, and 3) writing in the target language (English) in the Japanese source text.

### 3.6 Subtitle segments to be pre-edited

In order to evaluate quality improvement after application of the three pre-editing rules, the experimenters first pre-processed the existing TED subtitles by adjusting their alignments between the transcribed segments (Japanese) and human-translated ones (English) to correspond correctly .

Secondly, the adjusted segments were investigated to determine what types of pre-editing rules were needed according to the set of rules established in 3.5. Then we applied the missing rules to each segment (i.e. pre-edited) to make sure the segments satisfied all three elements compensated by the pre-editing rules. Table 2 summarizes the number of segments and which rules have been applied to them. Then we selected an equal number of segments from each 'Rules Application' category for quality evaluation, minimizing biased sampling of categories where different rules were applied.

| **Rules Application** | **Num. of Segments** |
|---|---|
| Rule 1 + 2 | 80 |
| Rule 1 + 3 | 19 |
| Rule 2 + 3 | 9 |
| Rule 1 + 2 + 3 | 5 |
| Total | 113 |

Table 2. Application of pre-editing rules

## 4 Evaluation methods

The effectiveness of the pre-editing rules was measured in terms of improvement of MT output quality given the 21-character-per-second (CPS) limitation. Since the translation target is TED subtitling, character limitation also needs to be taken into account.

### 4.1 Translation quality evaluation

Quality evaluation of MT outputs of both 'raw source' and 'pre-edited source' was carried out by human evaluators following the same guidelines. Along with it, we have also used an automatic evaluation, BLEU score, to investigate the correlation between human evaluation and BLEU.

Human evaluation was conducted by a Japanese speaker following the evaluation criteria shown in Table 3. The criteria were modified from a five-grade scale commonly used for MT system evaluation (Goto et al. 2013; Miyata & Fujita 2017). The reasons for employing the criteria in this study were 1) to minimize variations between human evaluators, and 2) to optimize for non-native English speakers.

The raters evaluated each segment using a three-point scale, with 3 indicating 'Good' and 1 indicating 'Nonsense.' For details, see Appendix A.

| Criterion | Score |
|---|---|
| Good | 3 |
| Acceptable | 2 |
| Nonsense | 1 |

Table 3. Human Evaluation Criteria

Mean score of total subtitle segments are calculated and compared between MT output of the raw source and the pre-edited source for improved quality.

BLEU score was also employed to evaluate the NMT outputs of the raw and pre-edited source texts against TED human translation as a reference text.

### 4.2 Inter-rater agreement

Prior to human evaluation, the inter-rater reliability (κ = 0.639) was confirmed to be within the range of "substantial agreement" (Landis and Koch, 1977). This attests to the reliability of the quality evaluation scale.

## 5 Results of evaluation

The evaluation results confirmed that, compared to the results of MT output of the raw source text (hereafter referred to as Raw MT), the MT output of the pre-edited source text (hereafter, Pre-Edit MT) made quality improvement in the average score of both human evaluation and BLEU. It is also revealed the total number of subtitle segments that resulted in score increase to be 41%. Although some score decreases were found in the pre-edited MT, most of the segments stayed above the 'Acceptable' level on the human evaluation scale.

In addition to translation quality, we have also examined the subtitling character limitation and verified that the number of segments in both raw MT and pre-edited MT output that violate the 21-CPS rule guideline by TED was almost none. Hence, it is concluded that pre-editing with the three rules does not preclude meeting the 21-CPS requirement.

The following sections show detailed results of each aspect.

### 5.1 Characters Per Second

This section touches on whether MT can translate the pre-edited source segment in accordance with the 21-CPS limit for TED subtitles. We calculated the number of characters used in each segment and the use ratio – the actual number of characters used in the segment divided by the maximum allowable characters.

The result reveals the number of segments in the pre-edited MT that violate the 21-CPS requirement to be just one segment. The average CPS in the pre-edit MT (12.5 CPS) has increased from that of the human translation (11.6); however, the difference is not statistically significant ($p > 0.01$ in Wilcoxon signed-rank test). Hence, it is concluded that pre-edited MT subtitles would meet the 21-CPS requirement of TED.

| Data Set | Avg. CPS | 21 CPS Violation |
|---|---|---|
| **Human translation** | 85.7 | 0 |
| **Raw MT** | 90.7 | 2 |
| **Pre-edited MT** | 90.2 | 1 |

Table 4: Average CPS and CPS violation

**Overall translation quality**

The average score of the human evaluation showed that the raw MT and the pre-edited

MT has a statistically significant difference in their translation quality.

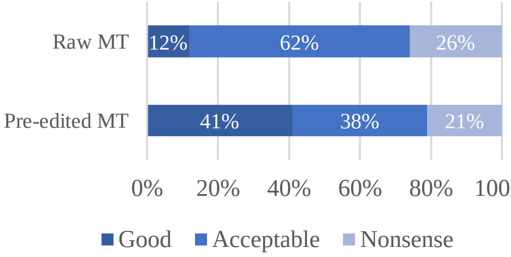
Table 5 shows that the raw MT output scored on average 1.85 on human evaluation and 7.70 on BLEU, which means these subtitles are, on average, ‘Acceptable,’ a translation functioning as adequate information with audiovisual elements.

In contrast, pre-edited MT output scored 2.21 on human evaluation and 9.32 on BLEU. The improvement of 0.36 from the raw MT on human evaluation is statistically significant ($p < 0.01$ in Wilcoxon signed-rank test).

| Data Set | Avg. Human Evaluation Score | BLEU |
|---|---|---|
| **Raw MT** | 1.8 | 7.70 |
| **Pre-edited MT** | 2.21 | 9.32 |
| **Difference** | 0.36* | 1.62 |

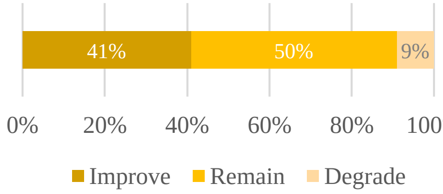
Table 5: Evaluation average score of raw and pre-edited MT

Graph 1 shows the percentages of quality levels, such as ‘Good,’ ‘Acceptable,’ and ‘Nonsense,’ of raw and pre-edited MT segments. It is notable that the number of pre-edited MT segments evaluated to be ‘Good’ increased from 12% to 41%.



Graph 1: Quality breakdown

Graph 2 below shows the percentages of segments that changed in quality or were unaffected after pre-editing. While half of the segments maintained the same quality rating, 41% of the pre-edited MT segments were improved and only 9%, were rated of lower quality.



Graph 2: Percentage of quality improvement, remain, and degrade

## 5.2 Examples of pre-edited NMT output

Pre-edited segments that improved in quality

Below is an example of a segment that improved on human evaluation from ‘Nonsense’ (1) to ‘Good’ (3). Adding punctuation and subjects to each sentence—simple rules—has made the quality of the subtitles much better.

| **ST / HT** |
|---|
| 上司が苦しい データが苦しい 考えるのが嫌になってしまった<br>Your boss was being difficult. The data was difficult. You become sick of thinking. |
| **Raw NMT (HE Score)** |
| I don't like the difficulty of my boss's difficult data. (1) |
| **PrE (HE Score)** |
| 上司が苦しい。データが苦しい。私は考えるのが嫌になってしまった。<br>My boss is difficult. I have difficulty in data. I hate to think about it. (3) |

Table 5. Example of HE Increase from Nonsense to Good

### 5.2.1 Pre-edited segments that degraded.

Although there is only a small number of segments that degraded after pre-editing, the following example below dropped two points on the scale from ‘Good’ to ‘Nonsense’.

| **ST / HT** |
|---|
| で思いもよらないアイデアが出てくる<br>You can come up with ideas that you wouldn't have thought of otherwise. |
| **Raw NMT (HE Score)** |
| There's an unexpected idea. (3) |
| **PrE (HE Score)** |
| で、自分の思いもよらないアイデアが出てくる。<br>So there's an idea that I don't want to think about. (1) |

Table 6. Example of degrading

The insertion of punctuation and subject was incorporated into this segment. A subject of the sentence, “自分の” (one’s own), was complemented, but how it was added was not sufficient, resulting in a nonsense translation.

If it “自分の” is replaced with “自分が” (“I” in the subject of a sentence), then the MT result improves, as shown below.

で、自分が思いもよらないアイデアが出てくる。
I have an idea that I can't think of.

## 6 Discussion

### 6.1 What are effective pre-editing rules for TED subtitling?

A set of pre-editing rules for TED subtitling intended for non-language expert use—insertion of punctuation, adding explicit subjects and objects, and writing proper nouns in the target language—was tested for its effect in this study. It was proven overall effective, with approximately 40% of the subtitle segments pre-edited with at least two of the rules reaching a ‘Good’ quality translation, although some lessening or lack of improvement in quality was also observed.

However, for practical use, implementation of these pre-editing rules in TED subtitling is, we feel, effective to improve overall readability. In addition, it is not yet clear what percentage of satisfactory MT outputs is needed to make potential audiences understand with additional audiovisual information, which may be a topic of our further research.

### 6.2 Does pre-editing affect the ‘readability’ of subtitles?

The readability of subtitles is another essential aspect of translation quality as regulated by the CPS rules. The result of this experiment shows pre-edited MT outputs meet the character limit requirement. Thus, for TED subtitle translation, the pre-editing rules and the pre-editing method can be effectively employed in this respect.

### 6.3 What skills are required for pre-editing in subtitling?

The editing rules were developed to be as simple as possible in order to enable monolingual speakers to perform pre-editing and disseminate their content in neural machine-translated text. However, our results could not rule out the possibility that editing performance may vary depending on pre-editor skill or knowledge. Further investigation into variants of editing rules for different pre-editors is therefore needed, including issues as to whether training may reduce user variation. Effective intralingual subtitling is simple and well-organized rather than ones that transcribe all speech including some fillers and misstatements.

## 7 Conclusion and further research

In this study, the authors developed a set of pre-editing rules for TED Talk subtitling to translate Japanese source text into English. The simplified rules optimized for @TexTra NMT were intended for use by monolingual pre-editors who can perform pre-editing for dissemination in English. This study investigated the effectiveness of the rules and confirmed quality improvements as evaluated by human ratings and BLEU score. The difference between raw MT and pre-edited MT output was statistically significant. However, there were some cases where pre-editing MT quality did not improve or even worsened the final product. In addition, variations in pre-editing were also confirmed, which may cause additional quality losses depending on the skill of the pre-editor. Lastly, the rules examined in this study were only for Textra NMT, so their effectiveness would need to be verified for use with other NMT systems, though we believe improvements would be leveraged.

## References

Common Sense Advisory. 2018. Machine Translation for Human Innovation. Retrieved on November 17, 2018: http://www.commonsenseadvisory.com/machine_translation.aspx

Goto, Isao, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proceedings of the 10th NII Testbeds and Community for Information access Research Conference*, pages 260-286.

Hiraoka, Yusuke and Masaru Yamada. 2019. Is Neural Machine Translation Capable of Subtitling?: Pre-editing Rules for Subtitling of TED Talks. In *Proceedings of the 25th Natural Language Processing conference*, pages 934-937.

Landis, Richard and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics,* 33(1): 159-174.

Miyata, Rei and Atsushi Fujita. 2017. Investigating the Effectiveness of Pre-Editing Strategy and the Diversity of Pre-Edit Operations for Better Use of Machine Translation. *Invitation to Interpreting and Translation Studies*, 18: 53-72.

## Appendix

### A. Human Evaluation Criteria

| | |
|---|---|
| Good | |
| 5 | Information of the original text has been completely translated. There are no grammatical errors in the translation. Word choice and phrasing are natural even from a native speaker's point-of-view. |
| 4 | Word choice and phrasing are slightly unnatural, but information of the original text has been completely translated and there are no grammatical errors in the translation. |
| Acceptable | |
| 3 | There are some minor translation errors with less important information, but the meaning of the original text can be easily understood. |
| 2 | Important parts of the original text are omitted or incorrectly translated, but the core meaning of the original text can still be understood with some effort. |
| Nonsense | |
| 1 | The meaning of the original text is incomprehensible. |

### B. TEDxTokyo Videos

TED presentations used for the present study were selected from the 2012 and 2013 TEDxTokyo. They are all categorized as a topic of "Business" on the TEDxTokyo website (https://www.tedxtokyo.com/).

| **Title (Japanese)** | | | **Num. of Segments (original/aligned)** | |
|---|---|---|---|---|
| **YouTube URL** | **Length** (mm:ss) | **Presenter** | **JPN** | **EN** |
| The treasure islands of Japan (Nihon no ritō ha takarajima)<br>https://www.youtube.com/watch?v=W_SBR3p_qyA | 8:49 | Isamoto, Atsuko | 128/61 | 130/61 |

| | | | | |
|---|---|---|---|---|
| Life balance (Raifu baransu)<br>https://www.youtube.com/watch?v=sd6OLoQW0hY | 12:14 | Komuro, Yoshie | 171/103 | 233/103 |
| Changing the world with spider webs (Kumo no ito de kawaru sekai)<br>https://www.youtube.com/watch?v=ldybnuFxdiQ | 5:54 | Sekiyama, Kazuhide | 199/60 | 148/60 |
| Play this word game to come up with original ideas (Atarashī aidea no tsukurikata)<br>https://www.youtube.com/watch?v=jzDwcNliXV8 | 8:41 | Takahashi, Shinpei | 108/71 | 105/71 |

## C. Examples of Pre-editing

Examples of the three pre-editing rules developed in this study are illustrated below.

**Punctuation insertion**

**Examples:**

Original text

今 日本は 少子化だけじゃなくうつ病の問題 ダイバーシティ 大介護の問題 財政難 問題山積の国です。
[Back Translation: The birth rate is not the only problem we're facing. All sorts of problems such as depression diversity elderly care financial problems are piled up.]

Pre-edited text

今、 日本は、 少子化だけじゃなく、 うつ病の問題、 ダイバーシティ、 大介護の問題、 財政難、 問題山積の国です。
[Back Translation: The birth rate is not the only problem we're facing. All sorts of problems, such as depression, diversity, elderly care, financial problems are piled up.]

**Note:**

No clear-cut rules are available for inserting punctuation in Japanese, but the way they are added above is to clarify the word boundaries to improve machine-translatability as well as human readability, since the Japanese writing system does not require spaces between words and sometimes word boundaries are ambiguous. Therefore, inserting punctuations such as commas, performed by pre-editors, would support MT quality improvement.

**Subject/Object insertion**

**Examples:**

Original text

だから会議が長引き 貧困なアイディアが出て 売れない 帰れない。
[Back Translation: So the meeting drags on and only poor ideas come up; won't sell; can't go home;]

Pre-edited text

だから会議が長引き 貧困なアイディアが出て 商品が売れない 社員は帰れない。
[Back Translation: So the meeting drags on and only poor ideas come up; the products won't sell; the employees can't go home;]

**Note:**

- It is necessary to add explicit subjects and/or objects of a sentence since the Japanese language is a pro-drop language in which certain pronouns are omitted when they are pragmatically or grammatically inferable.
- A sentence with a verb (predicate) requires a subject and object if applicable, so they have to be added by the pre-editor.
- Insertion of a subject "I" is, for most cases, not mandatory because it is often added

automatically in neural machine translation; however, it is still recommended to clarify the subject.

**Proper Noun**

**Examples:**

Original text

後５年で 日本の団塊世代は 一斉に70代に入ります。
[Back Translation: The baby boomers will be in their 70s in the next 5 years.]

Pre-edited text

後５年で 日本の The baby boomers は 一斉に70代に入ります。

**Note:**

- Machine translation is not yet good at translating proper nouns. Thus, when a proper noun is included in the original source text, one can either translate it into the target language (English) and write it in the source text or romanize it in the source text.