Understanding Pre-Editing for Black-Box Neural Machine Translation

Rei Miyata[†]   **Atsushi Fujita**[‡]

[†]Nagoya University, Nagoya, Japan
[‡]National Institute of Information and Communications Technology, Kyoto, Japan
`miyata@nuee.nagoya-u.ac.jp, atsushi.fujita@nict.go.jp`

Abstract

Pre-editing is the process of modifying the source text (ST) so that it can be translated by machine translation (MT) in a better quality. Despite the unpredictability of black-box neural MT (NMT), pre-editing has been deployed in various practical MT use cases. Although many studies have demonstrated the effectiveness of pre-editing methods for particular settings, thus far, a deep understanding of what pre-editing is and how it works for black-box NMT is lacking. To elicit such understanding, we extensively investigated human pre-editing practices. We first implemented a protocol to incrementally record the minimum edits for each ST and collected 6,652 instances of pre-editing across three translation directions, two MT systems, and four text domains. We then analysed the instances from three perspectives: the characteristics of the pre-edited ST, the diversity of pre-editing operations, and the impact of the pre-editing operations on NMT outputs. Our findings include the following: (1) enhancing the explicitness of the meaning of an ST and its syntactic structure is more important for obtaining better translations than making the ST shorter and simpler, and (2) although the impact of pre-editing on NMT is generally unpredictable, there are some tendencies of changes in the NMT outputs depending on the editing operation types.

1   Introduction

Recent advances in machine translation (MT) have greatly facilitated its practical use in various settings from business documentation to personal communication. In many practical cases, MT systems are used as black-box and one well-tested approach to make use of a black-box MT is pre-editing, i.e., modifying the source text (ST) to make it suitable for the intended MT system.

The effectiveness of pre-editing has so far been demonstrated in many studies (Pym, 1990; O'Brien and Roturier, 2007; Seretan et al., 2014). A study focusing on statistical MT (SMT) has also shown that more than 90% of an ST can be rewritten into a text that can be machine-translated with sufficient quality (Miyata and Fujita, 2017), exhibiting the potential of the pre-editing approach.

However, the feasibility and possibility of pre-editing for neural MT (NMT) has not been examined extensively. While efforts have recently been invested in the implementation of pre-editing strategies for black-box NMT settings, achieving improved MT quality (e.g., Hiraoka and Yamada, 2019; Mehta et al., 2020), the potential gains of pre-editing remain unexplored. Notably, the impact of pre-editing on black-box MT is unpredictable in nature. In particular, NMT models trained in an end-to-end manner can be sensitive to minor modifications of the ST (Cheng et al., 2019), which may affect the feasibility of pre-editing.

In short, while pre-editing has been implemented in practical MT use cases, what pre-editing is and how it works with black-box NMT systems remain open questions. To explore the possibility of pre-editing and its automation, in this study, we provide fine-grained analyses of human pre-editing practices and their impact on NMT. We systematically collected pre-editing instances in various conditions, i.e., translation directions, NMT systems, and text domains (§3). We then conducted in-depth analyses of the collected instances from the following three perspectives: the characteristics of the pre-edited ST (§4), the diversity of pre-editing operations (§5), and the impact of pre-editing operations on the NMT outputs (§6). The findings of these analyses provide useful insights into the effective and efficient implementation of pre-editing for the better use of black-box NMT systems in the future, as well as the robustness of current NMT systems when STs are manually perturbed.

2 Related Work

Pre-editing is the process of rewriting the source text (ST) to be translated in order to obtain better translations by MT. Though the scope of effective pre-editing operations depends on the downstream MT system and there is no deterministic relation between pre-editing operations and the quality of MT output, its effectiveness has been demonstrated for various translation directions, MT architectures, and text domains.

Manual pre-editing has long been implemented in combination with controlled languages (Pym, 1990; Reuther, 2003; Nyberg et al., 2003; Kuhn, 2014). In the period of rule-based MT (RBMT), pre-editing was considered as a promising approach since the behaviour of RBMT is more predictable and controllable. For example, O'Brien and Roturier (2007) examined the impact of English controlled language rules on two different MT engines, revealing the rules of high effectiveness. The pre-editing approach with controlled languages has also been tested for statistical MT (SMT) (Aikawa et al., 2007; Hartley et al., 2012; Seretan et al., 2014). These studies developed or utilised a set of controlled language rules for rewriting ST. While these rule sets are optimised for particular MT systems and differ from each other, we can observe some shared characteristics among them. In particular, rules that prohibit long sentences (e.g., of more than 25 words) are widely adopted in the existing rule sets (O'Brien, 2003).

Automation of pre-editing is also an important research field in natural language processing. Semi-automatic tools such as controlled language checkers (Bernth and Gdaniec, 2001; Mitamura et al., 2003) and interactive rewriting assistants (Mirkin et al., 2013; Gulati et al., 2015) were developed to facilitate manual pre-editing activities. Fully automatic pre-editing has long been explored (e.g., Shirai et al., 1998; Mitamura and Nyberg, 2001; Yoshimi, 2001; Sun et al., 2010). In particular, many researchers have examined methods of *reordering* the source-side word order as a pre-translation processing (Xia and McCord, 2004; Li et al., 2007; Hoshino et al., 2015). While the re-ordering approach has generally proven effective for SMT, its effectiveness for NMT is not obvious; negative effects have even be reported (Zhu, 2015; Du and Way, 2017). In recent years, techniques of automatic text simplification have been applied to improve NMT outputs (Štajner and Popović, 2018;

Mehta et al., 2020). The underlying assumption of these studies is that simpler sentences are more machine translatable.

Previous studies have investigated various pre-editing methods from different perspectives, focusing on different linguistic phenomena. Indeed, individual research has led to improved MT results. However, what is crucially needed is a broad understanding of what pre-editing is and how it works. For example, Miyata and Fujita (2017) addressed this issue by collecting instances of *bilingual pre-editing*, i.e., pre-editing ST while referring to its MT output, done by human editors and analysing them in detail. They demonstrated the maximum gain of pre-editing for an SMT and provided a comprehensive typology of editing operations. Nevertheless, their study has two major limitations: (1) recent NMT was not examined, and (2) practical insights for better practices of pre-editing were not sufficiently presented.

NMT models trained in an end-to-end manner behave very differently from SMT and RBMT, which, in turn, affects pre-editing practices. As reported in several studies, despite their rapid improvement, NMT models are still vulnerable to input noise (Belinkov and Bisk, 2018; Ebrahimi et al., 2018; Cheng et al., 2019; Niu et al., 2020). The pre-editing operations identified in previous studies are not necessarily effective for current black-box NMT systems.[1] For example, Marzouk and Hansen-Schirra (2019) adopted nine controlled language rules[2] and evaluated their impact on the MT output for German-to-English translation in the technical domain. The human evaluation results revealed that these rules improved the performance of the RBMT, SMT, and hybrid systems, but did not have positive effects on the NMT system. Hiraoka and Yamada (2019) demonstrated the effectiveness of the following three pre-editing rules in improving Japanese-to-English TED Talk subtitle

---

[1] The ideal goal of the pre-editing approach is to adapt the STs to what the intended NMT system can properly translate, and in the end, what it has been trained on, i.e., training data. For a black-box MT system, because we cannot directly refer to its training data, we should grasp its statistical characteristics indirectly through MT output.

[2] The rules are as follows: (1) using straight quotes for interface texts, (2) avoiding light-verb construction, (3) formulating conditions as if sentences, (4) using unambiguous pronominal references, (5) avoiding participial constructions, (6) avoiding passives, (7) avoiding constructions with "sein" + "zu" + infinitive, (8) avoiding superfluous prefixes, and (9) avoiding omitting parts of the words (Marzouk and Hansen-Schirra, 2019, p.184).

5. Perfect	Information in the original text has been completely translated. There are no grammatical errors in the translation. The word choice and phrasing are natural even from a native speaker's point of view.
4. Good	The word choice and phrasing are slightly unnatural, but the information in the original text has been completely translated, and there are no grammatical errors in the translation.
3. Fair	There are some minor errors in the translation of less important information in the original text, but the meaning of the original text can be easily understood.
2. Acceptable	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.
1. Incorrect/nonsense	The meaning of the original text is incomprehensible.

Table 1: MT evaluation criterion adopted in Miyata and Fujita (2017): The "Perfect" and "Good" ratings are regarded as satisfactory quality.
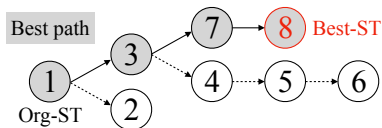

Figure 1: Tree representation of ST versions in a unit.

Name	Domain	Mode	Size	Avg. length (S.D.)
hospital	hospital conversation	spoken	25	13.0 (4.7)
municipal	municipal procedure	written	25	20.4 (10.7)
bccwj	Japanese-origin news article from BCCWJ	written	25	28.6 (18.6)
reuters	English-origin news article from Reuters	written	25	36.8 (15.3)

Table 2: Statistics for the Org-ST datasets for pre-editing.

translation using a black-box NMT system: (1) inserting punctuation, (2) making implied subjects and objects explicit, and (3) writing proper nouns in the target language (English).

As these studies cover a limited range of linguistic phenomena, translation directions, and text domains, we are not in the position to draw decisive conclusions; we still do not know what types of pre-editing operations are possible and how NMT is affected when these operations are performed. To elicit the best pre-editing practices for NMT, as a starting point, we need to understand what is happening and what can be obtained in the process of pre-editing, while also re-examining the previous findings and conventional methods.

3 Collection of Pre-Editing Instances

3.1 Protocol

To collect fine-grained manual pre-editing instances, we adopted the protocol formalised by Miyata and Fujita (2017), in which a human editor incrementally and minimally rewrites an ST on a trial-and-error basis with the aim of obtaining better MT output. An original ST (Org-ST) and its pre-edited versions are collectively called a *unit*. Using an online editing platform we developed, editors implement the protocol in the following steps:

Step 1. Evaluate the MT output of the current ST based on a 5-point scale criterion shown in Table 1. If the quality of the MT output is satisfactory (i.e., "Perfect" or "Good"), go to Step 4; otherwise, go to Step 2.

Step 2. Select one of the versions of the ST in the unit to be rewritten and go to Step 3. If none of the versions are likely to become satisfactory through further edits, go to Step 4.

Step 3. Minimally edit the ST[3] while maintaining its meaning, referring to the corresponding MT output. The MT output for the edited ST is automatically generated and registered in the unit. Return to Step 1.

Step 4. Select one version of the ST that achieves the best MT quality (Best-ST) from among all the versions in the unit, and terminate the process for the unit.

The pre-editing instances in a unit collected through this protocol form a tree structure as shown in Figure 1. We refer to the shortest path between the Org-ST and the Best-ST as *Best path*. An important extension to the work in Miyata and Fujita (2017) is that our platform provides editors with a visualisation of the tree representation of the pre-editing history. This can facilitate the selection of ST versions in Step 2.

3.2 Implementation

To extensively investigate pre-editing phenomena, we prepared the following conditions:

Translation directions: We targeted Japanese-to-English (Ja-En), Japanese-to-Chinese (Ja-Zh), and Japanese-to-Korean (Ja-Ko) translations.

---

[3]We operationally defined "to minimally edit" as "to modify an ST with a small edit that is difficult to be further decomposed into more than one independent edit, without inducing ungrammaticality in the edited sentence."

Lang.	System	Domain	Num. of pre-editing instances				Num. of units	
			Total	Avg.	Med.	Max	Org=Satisfactory	Best=Satisfactory
Ja-En	Google	hospital	255	10.2	7	55	4/25	**25/25**
		municipal	162	6.5	5	44	9/25	**25/25**
		bccwj	545	21.8	10.5	171	7/25	23/25
		reuters	370	14.8	6.5	80	7/25	**25/25**
	TexTra	hospital	139	5.6	5.5	25	7/25	**25/25**
		municipal	136	5.4	4	35	10/25	**25/25**
		bccwj	493	19.7	11.5	79	2/25	22/25
		reuters	492	19.7	18	86	4/25	24/25
Ja-Zh	Google	hospital	264	10.6	10	30	0/25	24/25
		municipal	376	15.0	13	41	0/25	23/25
		bccwj	427	17.1	16	41	2/25	20/25
		reuters	304	12.2	10	27	0/25	24/25
	TexTra	hospital	160	6.4	6.5	15	1/25	**25/25**
		municipal	172	6.9	7	20	2/25	**25/25**
		bccwj	231	9.2	5	38	4/25	22/25
		reuters	249	10.0	7	31	1/25	22/25
Ja-Ko	Google	hospital	209	8.4	9	22	0/25	**25/25**
		municipal	225	9.0	8	26	0/25	**25/25**
		bccwj	223	8.9	7	27	1/25	22/25
		reuters	293	11.7	10	33	0/25	24/25
	TexTra	hospital	160	6.4	6	26	2/25	**25/25**
		municipal	171	6.8	5	32	2/25	**25/25**
		bccwj	277	11.1	6	28	3/25	23/25
		reuters	319	12.8	11	38	1/25	23/25

Table 3: Statistics for the collected pre-editing instances and the MT quality achievement.

MT systems: As black-box MT systems, we adopted Google Translate[4] and TexTra.[5] Both are general-purpose NMT systems that are prevalently used for translating Japanese texts into other languages.

Text domains: We selected four text domains, whose linguistic characteristics, such as mode and sentence length, are different from each other (see Table 2 for details).

We randomly selected 25 Japanese sentences for each of the four text domains, and used the resulting ST set consisting of 100 sentences for all of the six combinations of translation direction and MT system. We assigned one editor to each translation direction. Each editor was asked to work with both MT systems, without being informed of the type of MT system used in the task. All editors were professional translators with sufficient writing skills in Japanese and experience for evaluating MT outputs. Before the commencement of the formal tasks, we trained the editors using example sentences so that they could become accustomed to the task and platform.

The Ja-En task was implemented from November to December 2019; the Ja-Zh and Ja-Ko tasks were implemented from December 2019 to February 2020.

3.3 Statistics

Table 3 shows statistics for the pre-editing instances collected through the protocol described above. In general, the numbers of collected instances for the **hospital** and **municipal** domains were smaller than those for the **bccwj** and **reuters** domains, reflecting the influence of sentence length of the Org-ST. In other words, the shorter the sentence is, the fewer parts there are to be edited.

A notable finding is that while only about 11% (69/600) of the MT output for the Org-ST was of satisfactory quality, 95% (571/600) of the MT output of the Best-ST was satisfactory. This means that almost all the ST can be pre-edited into a form that can lead to satisfactory MT output, demonstrating the potential of both pre-editing and NMT.

The number of collected instances can be interpreted as the editing efforts required to obtain the Best-ST from the Org-ST. In most of the settings, the median number of collected instances for a unit falls in the range of 5 to 10. It is thus necessary to optimise the pre-editing process for an intended MT system. The length of the Best path approximates the minimum editing efforts needed to obtain the Best-ST. The total number of pre-editing instances in the Best path was 2,443, while the total of all instances is 6,652. This implies that there is substantial opportunity for reduction of the pre-editing efforts.

		Org-ST	Ja-En (Best-ST)		Ja-Zh (Best-ST)		Ja-Ko (Best-ST)	
			Google	TexTra	Google	TexTra	Google	TexTra
Sentence length	**Avg.**	25.4	27.8	26.9	28.6	27.1	27.8	26.9
	S.D.	16.3	17.6	16.7	17.2	16.0	16.7	16.6
	Med.	19.5	21.5	20	23	22	22.5	20.5
Attachment distance	**Avg.**	1.95	1.97	1.99	1.99	1.99	2.00	1.98
(Avg. per sentence)	**S.D.**	0.65	0.53	0.65	0.60	0.63	0.64	0.62
	Med.	1.83	2.00	1.96	2.00	1.98	2.00	1.91
Dependency depth	**Avg.**	3.57	3.73	3.68	3.73	3.77	3.78	3.76
	S.D.	1.91	1.97	1.88	1.89	1.93	1.85	1.92
	Med.	3	3	3	3	4	4	4
Lexical diversity	**Token (A)**	2,538	2,779	2,685	2,861	2,709	2,780	2,693
	Type (B)	1,010	1,074	1,060	1,106	1,061	1,068	1,055
	A/B	2.513	2.588	2.533	2.587	2.553	2.603	2.553
Word frequency rank	**25th**	7	7	7	7	7	7	7
(Percentile)	**50th (Med.)**	170	143	154	143	155	143	169.5
	75th	2655	2304.25	2458	2471	2554	2470	2593.5

Table 4: Linguistic characteristics of the Org-ST and Best-ST.

4 Characteristics of Pre-Edited Sentences

To understand the differences between the original and pre-edited STs, in this section, we describe their general linguistic characteristics. Here, we compare the Org-ST and the Best-ST that achieved a satisfactory MT result in order to elicit the features of machine translatable ST.

4.1 Structural Characteristics

To quantify structural complexity, we used the following three indices:

(1) sentence length: the number of words per sentence[6]

(2) attachment distance: the averaged distance of all attachment pairs of the Japanese base phrases in a sentence

(3) dependency depth: the maximum distance from the root word in the dependency tree

We used the Japanese tokeniser MeCab[7] to calculate (1) and the Japanese dependency parser JUMAN/KNP[8] to calculate (2) and (3).

The first three blocks in Table 4 show the results for these indices. It is evident on all indices, the Org-ST exhibits the lowest scores. In other words, the length and surface complexity of the sentences generally increased through the pre-editing operations. This is a counter-intuitive finding in that most previous pre-editing practices have axiomatically assumed that shorter and less complex sentences are better for MT. We further delve into this in §5.

4.2 Lexical Characteristics

The remaining two blocks in Table 4 present statistics for the lexical characteristics of the STs. The results for lexical diversity indicate that both the total number of word types and the Token/Type ratio increased from the Org-ST to the Best-ST for all the conditions. This suggests that though the diversity of words increased slightly, the word distribution became peakier through pre-editing.

We also calculated the word frequency rank with Wikipedia as the reference.[9] To assess the status of word frequency in relation to MT, it would be ideal to use the training data for each MT system, but such data are unavailable in black-box MT settings. Therefore, we decided to use Wikipedia as a convenient way to observe general word frequency. Lower numbers indicate higher word frequencies in Wikipedia. The 50th and 75th percentile values in the datasets imply that pre-editing induced the avoidance of low-frequency words.

To further inspect the differences between the Org-ST and the Best-ST, we extracted the word types (a) that appeared only in the Org-ST and (b) that appeared only in the Best-ST. Figure 2 illustrates the rank distributions of (a) and (b) for each condition. It is clear that low-frequency words with a frequency rank of around 10,000 decreased in the Best-ST, while words with a frequency rank of around 2,000–4,000 increased in the Best-ST. As Koehn and Knowles (2017) demonstrated, low-frequency words still pose major obstacles for NMT systems. Our results endorse this claim from a different perspective and can provide general strategies for word choice in the pre-editing task.

---

[6]If ST instance includes multiple sentences, we averaged the scores.

[7]https://taku910.github.io/mecab/

[8]http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[9]We used the whole text data of Japanese Wikipedia obtained in October 2019 (https://dumps.wikimedia.org/).
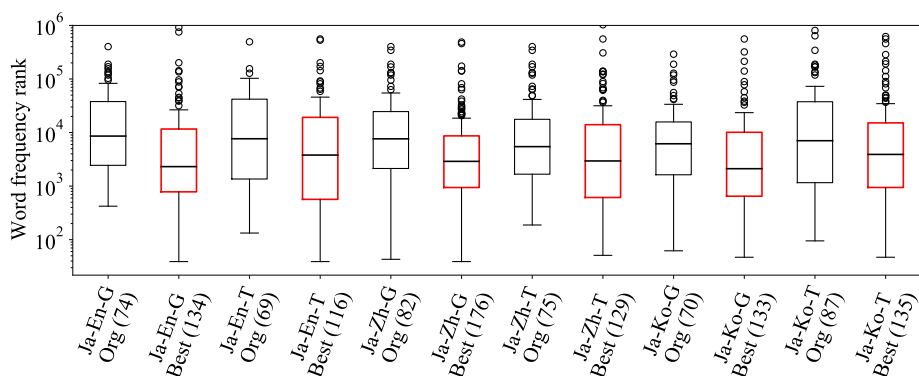
Figure 2: Differences in word frequency rank distribution between the Org-ST and Best-ST (G: Google, T: TexTra). The numbers in parentheses indicate the number of instances, i.e., word types.

5 Diversity of Pre-Editing Operations

5.1 Typology of Edit Operations

To understand the diversity of edit operations for pre-editing, we manually annotated the collected pre-editing instances in terms of linguistic operations. Given that the Best path contains effective editing operations for improved MT quality, we focused on the pairs of ST versions in the Best path (e.g., the pairs {1→3, 3→7, 7→8} in Figure 1). We randomly selected 10 units for each of the 24 combinations of translation direction, MT system, and text domain, resulting in a total of 961 pre-editing instances. We then excluded 26 instances that could be decomposed into multiple smaller edits[10] and classified the remaining 935 instances, each of which consists of a minimum edit of ST, based on the typology proposed by Miyata and Fujita (2017). Through the classification, we refined the existing typology to consistently accommodate all the instances.

Table 5 presents our typology of editing operations with the number of instances in the different conditions. The typology consists of 39 operation types under 6 major categories, which enables us to grasp the diversity and trends of pre-editing operations. Compared to structural editing, local modifications of words and phrases were frequently used in the Best path. The dominant type is **C01 (Use of synonymous words)**: content words are replaced by another synonymous word. This operation is important for achieving appropriate word choice in the MT output. **C07 (Change of content)**, the second dominant type, includes the ad-

dition of information that is inferred by human editors based on the intra-sentential context or even external knowledge. For example, a named entity '*Nemuro-sho*' (Nemuro office) was changed into '*Nemuro-keisatsu-sho*' (Nemuro police office) by using the knowledge of the entity. It might be challenging to automate such *creative* operations.

It is also notable that **S01 (Sentence splitting)** only amounts to 1.5% of all instances, which supports the observation in §4.1 that in general, sentence length was not reduced, and even increased by pre-editing. Among the 14 cases of this type, nine of the split sentences were 60–67 words in length. These results support the empirical observation by Koehn and Knowles (2017) that NMT systems still have difficulty in translating sentences longer than 60 words, and suggest that sentence splitting may only be promising for such very long sentences.

5.2 Strategies for Effective Pre-Editing

Towards the effective exercise of pre-editing, we further analysed the pre-editing instances in terms of *informational strategies* based on the notion of explicitation/implicitation acknowledged in translation studies (Vinay and Darbelnet, 1958; Chesterman, 1997; Murtisari, 2016). Following these studies, we broadly defined explicitation as an act of indicating what is implied in the text to clarify its meaning and implicitation as the inverse act of explicitation. We classified all the instances analysed above except for the **E01** and **E02** types into three general strategies, namely, explicitation, implicitation, and (information) preservation. The right side of Table 5 shows the classification result. The total numbers of instances classified into each strategy were 329, 88, and 480, respectively. Not surprisingly, this indicates that explicitation is an essential

---

[10] Only 2.7% of the edits were not regarded as minimum, which demonstrated satisfactory adherence to our instructions, compared with the implementation by Miyata and Fujita (2017), in which 568 pre-editing instances were finally decomposed into 979 instances.

ID	Editing operation type	Ja-En		Ja-Zh		Ja-Ko		Total	Expl.	Impl.	Pres.
		G	T	G	T	G	T				
S01	Sentence splitting	1	0	3	3	4	3	14	0	0	14
S02	Structural change	3	5	9	4	4	2	27	8	1	18
S03	Use/disuse of topicalisation	1	7	4	3	1	3	19	5	2	12
S04	Insertion of subject/object	2	1	1	3	5	2	14	14	0	0
S05	Use/disuse of clause-ending noun	3	2	2	2	2	1	12	12	0	0
S06	Change of voice	1	3	0	0	0	0	4	2	0	2
S07	Other structural changes	1	0	2	1	1	0	5	3	0	2
P01	Insertion/deletion of punctuation	19	16	5	12	9	10	71	0	0	71
P02	Use/disuse of chunking marker(s)	6	12	2	1	3	4	28	11	8	9
P03	Phrase reordering	6	4	7	1	9	4	31	0	0	31
P04	Change of modification	1	3	3	0	0	0	7	0	0	7
P05	Change of connective expression	3	18	4	2	10	3	40	24	5	11
P06	Change of parallel expression	3	8	2	8	4	11	36	7	2	27
P07	Change of apposition expression	1	7	2	1	1	4	16	8	4	4
P08	Change of noun/verb phrase	1	3	2	1	3	3	13	9	3	1
P09	Use/disuse of compound noun	1	5	2	2	6	12	28	16	12	0
P10	Use/disuse of affix	4	4	1	2	3	3	17	1	0	16
P11	Change of sahen noun expression	0	1	1	1	2	0	5	1	0	4
P12	Change of formal noun expression	1	2	2	2	2	0	9	4	0	5
P13	Other phrasal changes	0	1	0	1	2	1	5	4	0	1
C01	Use of synonymous words	18	18	19	18	25	20	118	14	10	94
C02	Use/disuse of abbreviation	2	7	2	2	1	7	21	19	2	0
C03	Use/disuse of anaphoric expression	4	4	2	2	1	1	14	10	2	2
C04	Use/disuse of emphatic expression	1	2	2	1	4	1	11	10	1	0
C05	Category indication/suppression	5	3	6	5	4	7	30	29	1	0
C06	Explanatory paraphrase	3	4	1	0	1	1	10	0	0	10
C07	Change of content	22	20	21	9	14	8	94	57	23	14
F01	Change of particle	9	14	4	6	7	7	47	13	5	29
F02	Change of compound particle	8	5	5	2	5	6	31	24	2	5
F03	Change of aspect	1	4	1	0	5	1	12	0	0	12
F04	Change of tense	0	0	1	1	1	1	4	0	0	4
F05	Change of modality	3	1	2	1	3	1	11	5	0	6
F06	Use/disuse of honorific expression	3	1	1	2	2	1	10	0	0	10
O01	Japanese orthographical change	10	16	9	5	9	12	61	12	4	45
O02	Change of half-/full-width character	0	5	3	2	2	4	16	7	1	8
O03	Insertion/deletion/change of symbol	0	2	0	0	0	0	2	0	0	2
O04	Other orthographical change	0	1	0	0	3	0	4	0	0	4
E01	Grammatical errors	0	8	5	2	2	5	22	–	–	–
E02	Content errors	5	0	8	1	1	1	16	–	–	–

Table 5: Constructed typology of editing operations (G: Google, T: TexTra). The first letter of ID indicates the six major categories (S: Structure, P: Phrase, C: Content word, F: Functional word, O: Orthography, E: Errors casually introduced in the ST). The right three columns provide the frequencies for general informational strategies (Expl.: Explicitation, Impl.: Implicitation, Pres.: Preservation).

strategy for effective pre-editing.

We also grouped all the 329 instances of explicitation into the following four subcategories.[11]

Information addition is the strategy of adding supplementary information, such as subjects, modality, and explanation, to clarify the content of the ST. For example, subjects were sometimes inserted as they tend to be omitted in Japanese sentences. This strategy generally corresponds to operation **C07 (Change of content)** described earlier.

Use of clear relation includes structural changes and the use of explicit connective markers to make the relation between words, phrases, and clauses more intelligible. For example, the relation between the subject and object can be clarified by using the nominative case marker 'ga' in Japanese.

Use of narrower sense is the strategy of replacing general words with more specific ones. For example, the verb '*dasu*,' which has multiple meanings such as 'put,' 'take,' and 'send,' was replaced with the verb '*teishutsusuru*,' which has a narrower range of meaning and was correctly translated as 'submit.'

Normalisation includes the use of authorised or standardised expressions, style, and notation. For example, elliptic sentence-ending was completed to construct a normal structure.

		Ja-En		Ja-Zh		Ja-Ko	
		Google	TexTra	Google	TexTra	Google	TexTra
TER	**Pearson's r**	0.244	0.217	0.144	0.204	0.580	0.347
	Spearman's ρ	0.218	0.184	0.094	0.172	0.574	0.248
Num. of edits	**Pearson's r**	0.264	0.153	0.205	0.181	0.465	0.212
	Spearman's ρ	0.210	0.221	0.219	0.226	0.449	0.245

Table 6: Correlation of the TER and the number of edits between ST and MT.

These strategies can be used as concise pre-editing principles for human editors and can guide researchers in devising effective tools for pre-editing. We also emphasise that these general informational strategies are not specific to the Japanese language and could be applied to other languages.

6 Impact of Pre-Editing on Neural Machine Translation

This section investigates how pre-editing operations affect the NMT output. As indicated in §2, NMT systems still lack robustness, and minor modifications of the input would drastically change the output. From the practical viewpoint of deploying pre-editing, predictability is an important object to pursue. Here, we examine the impacts of minimum edits of the ST on the NMT output. To measure the amount of text editing, hereafter, we use the Translation Edit Rate (TER), which is calculated by dividing the number of edits (insertion, deletion, substitution, and shift) required to change a string into the reference string by the average number of reference words (Snover et al., 2006). For any consecutive pair of STs or their corresponding MT outputs, we used the chronologically later version as the reference. For word-level tokenisation, we used MeCab for Japanese, NLTK[12] for English, jieba[13] for Chinese, and KoNLPy[14] for Korean.

6.1 Correlation of the Amount of Edits between the ST and MT

To grasp the general tendency, using all the collected pre-editing instances (see Table 3), we first calculated the correlation coefficients (Pearson's r and Spearman's ρ) between the amount of edits (the TER and the number of edits) in the ST and in the MT. More formally, let ST$'$ be the pre-edited versions of ST. For TER, the correlation is between TER(ST, ST$'$) and TER(MT(ST), MT(ST$'$)). For the number of edits, the

correlation is between EditCount(ST, ST$'$) and EditCount(MT(ST), MT(ST$'$)).

As shown in Table 6, most coefficients are in the range of 0.15–0.25, suggesting a very weak correlation. This means that the change in NMT output is hardly predictable based on the amount of edits in the ST. For example, the replacement of a single particle in the ST sometimes caused drastic changes of lexical choices in the MT output.

The Japanese-to-Korean translation is an exception; in particular, the correlation coefficients of the TER for the Google NMT system, i.e., 0.580 for Pearson's r and 0.574 for Spearman's ρ, indicate a moderate positive relationship between the changes in the ST and those in the MT. This is partly attributable to the fact that the syntactic structures of Japanese and Korean, including the word order and usage of particles, are substantially close. Thus, it is relatively easy to build sufficiently accurate MT systems.

6.2 Impact of Editing Operations on NMT

Finally, using the pre-editing instances in the Best path analysed in §5, we further investigated to what extent each type of minimum editing operation affects the MT output. At this stage, we focused on the 28 editing types that have at least 10 instances, considering that it is difficult to derive reliable insights from fewer data.

Figure 3 presents the distribution of the degree of changes in the MT output when an ST is pre-edited, measured by TER(MT(ST), MT(ST$'$)). Most of the structural edits (**S01–S04**) resulted in sizeable changes in the MT. This is reasonable since structural modifications in the ST tended to cause major changes in the MT as well, leading to high TER. In contrast, many of the editing types that include local modifications of functional words and orthographic notations (**F01–F03, F05, F06, O01, O02**) did not have major impacts on the MT results.

It is worth noticing that **P03 (Phrase reordering)** did not drastically affect the MT output. In other words, recent NMT systems in practical use manage to retain the phrase-level equivalence even

---

[12]https://www.nltk.org/index.html
[13]https://github.com/fxsjy/jieba
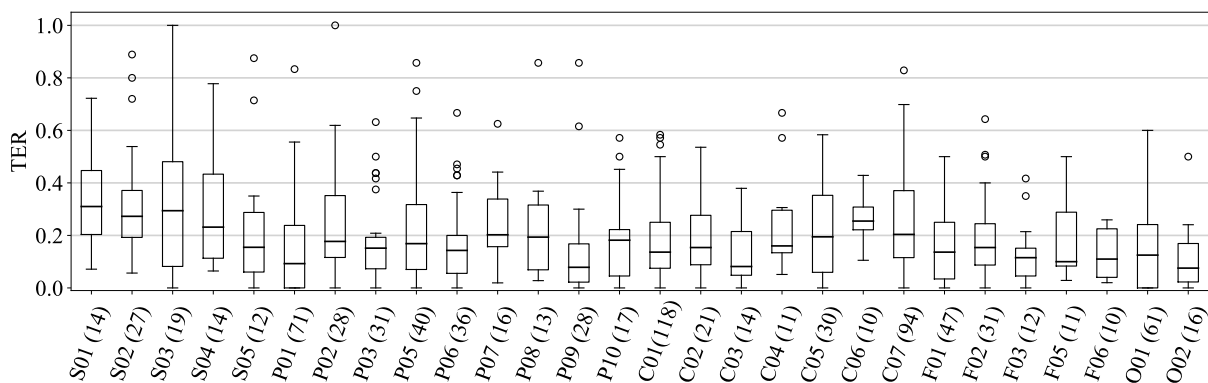[14]https://konlpy.org/en/latest/api/konlpy.tag/#module-konlpy.tag._kkma

Figure 3: Distribution of the TER for changes in the MT for each operation type with at least 10 instances. The numbers in parentheses indicate the number of instances.

when the position of a phrase is shifted. The influence of **P02 (Use/disuse of chunking marker(s))** is fairly significant. For human readers, the use of chunking markers, such as double quotes and square brackets, does not greatly affect the sentence parsing, but for NMT, it might seriously impinge on the tokenisation result, eventually leading to a large change in the final output.

7 Conclusion and Outlook

Towards a better understanding of pre-editing for black-box NMT settings, in this study, we collected instances of manual pre-editing in various conditions and conducted in-depth analyses of the instances. We implemented a human-in-the-loop protocol to incrementally record minimum edits of ST for all combinations of three translation directions, two NMT systems, and four text domains, and obtained a total of 6,652 instances of manual pre-editing. Since more than 95% of the STs were successfully pre-edited into one that led to a satisfactory MT quality, our collected instances contain empirical, tacit human knowledge on the effective use of black-box NMT systems. We also investigated the collected data from three perspectives: the characteristics of the pre-edited STs, the diversity of pre-editing operations, and the impact of pre-editing operations on the NMT output. The remarkable findings can be summarised as follows:

- Contrary to the acknowledged practices of pre-editing, the operation of making source sentences shorter and simpler was not frequently observed. Rather, it is more important to make the content, syntactic relations, and word senses clearer and more explicit, even if the ST becomes longer.

- As indicated by recent studies, the NMT systems are still sensitive to minor edits in the ST, and are unpredictable in general. However, there are recognisable tendencies in the MT output according to the types of editing operations, such as the relatively small impact of phrase reordering on NMT.

In future work, we plan to explore the effective implementation of pre-editing. The findings of this study provide a broad overview of the range of pre-editing operations and their expected benefits, which enables us to find feasible pre-editing solutions in practical use cases of black-box NMT systems. To develop automatic pre-editing tools using a collection of pre-editing instances, we need to handle the data insufficiency issue in machine learning, filling the gap between the training data and targeted black-box MT systems.

Moreover, as our pre-editing instances contain a wide variety of perturbations in the ST, they can also be used to evaluate the robustness of MT systems, which can lead to advances in MT research. We aim to jointly improve the two wheels of translation technology: pre-editing and MT.

References

Takako Aikawa, Lee Schwartz, Ronit King, Monica Corston-Oliver, and Carmen Lozano. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proceedings of the Machine Translation Summit XI*, pages 1–7, Copenhagen, Denmark.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, pages 1–13, Vancouver, Canada.

Arendse Bernth and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation*, 16(3):175–218.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4324–4333. Florence, Italy.

Andrew Chesterman. 1997. *Memes of Translation*. John Benjamins, Amsterdam.

Jinhua Du and Andy Way. 2017. Pre-reordering for neural machine translation: Helpful or harmful? *The Prague Bulletin of Mathematical Linguistics*, 108:171–182.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 653–663, Santa Fe, New Mexico, USA.

Asheesh Gulati, Pierrette Bouillon, Johanna Gerlach, Victoria Porro, and Violeta Seretan. 2015. The ACCEPT Academic Portal: A user-centred online platform for pre-editing and post-editing. In *Proceedings of the 7th International Conference of the Iberian Association of Translation and Interpreting Studies (AIETI)*, Malaga, Spain.

Anthony Hartley, Midori Tatsumi, Hitoshi Isahara, Kyo Kageura, and Rei Miyata. 2012. Readability and translatability judgments for 'Controlled Japanese'. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 237–244, Trento, Italy.

Yusuke Hiraoka and Masaru Yamada. 2019. Pre-editing plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks. In *Proceedings of the Machine Translation Summit XVII*, pages 64–72, Dublin, Ireland.

Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, Katsuhiko Hayashi, and Masaaki Nagata. 2015. Discriminative preordering meets Kendall's τ maximization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 139–144, Beijing, China.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the 1st Workshop on Neural Machine Translation (NMT)*, pages 28–39, Vancouver, Canada.

Tobias Kuhn. 2014. A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170.

Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 720–727, Prague, Czech Republic.

Shaimaa Marzouk and Silvia Hansen-Schirra. 2019. Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation*, 33(1-2):179–203.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. 2020. Simplify-then-translate: Automatic preprocessing for black-box translation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 8488–8495, New York, USA.

Shachar Mirkin, Sriram Venkatapathy, Marc Dymetman, and Ioan Calapodescu. 2013. SORT: An interactive source-rewriting tool for improved translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), System Demonstrations*, pages 85–90, Sofia, Bulgaria.

Teruko Mitamura, Kathryn L. Baker, Eric Nyberg, and David Svoboda. 2003. Diagnostics for interactive controlled language checking. In *Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT/CLAW)*, pages 237–244, Dublin, Ireland.

Teruko Mitamura and Eric Nyberg. 2001. Automatic rewriting for controlled language translation. In *Proceedings of the NLPRS2001 Workshop on Automatic Paraphrasing: Theories and Applications*, pages 1–12, Tokyo, Japan.

Rei Miyata and Atsushi Fujita. 2017. Dissecting human pre-editing toward better use of off-the-shelf machine translation systems. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 54–59, Prague, Czech Republic.

Elisabet Titik Murtisari. 2016. Explicitation in Translation Studies: The journey of an elusive concept. *The International Journal for Translation & Interpreting Research*, 8(2):64–81.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8538–8544, Online.

Eric Nyberg, Teruko Mitamura, and Willem-Olaf Huijsen. 2003. Controlled language for authoring and translation. In Harold Somers, editor, *Computers and Translation: A Translator's Guide*, pages 245–281. John Benjamins, Amsterdam.

Sharon O'Brien. 2003. Controlling controlled English: An analysis of several controlled language rule sets. In *Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT/CLAW)*, pages 105–114, Dublin, Ireland.

Sharon O'Brien and Johann Roturier. 2007. How portable are controlled language rules? In *Proceedings of the Machine Translation Summit XI*, pages 345–352, Copenhagen, Denmark.

Peter Pym. 1990. Pre-editing and the use of simplified writing for MT. In Pamela Mayorcas, editor, *Translating and the Computer 10: The Translation Environment 10 Years on*, pages 80–95. Aslib, London.

Ursula Reuther. 2003. Two in one – Can it work?: Readability and translatability by means of controlled language. In *Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (EAMT/CLAW)*, pages 124–132, Dublin, Ireland.

Violeta Seretan, Pierrette Bouillon, and Johanna Gerlach. 2014. A large-scale evaluation of pre-editing strategies for improving user-generated content translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 1793–1799, Reykjavik, Iceland.

Satoshi Shirai, Satoru Ikehara, Akio Yokoo, and Yoshifumi Ooyama. 1998. Automatic rewriting method for internal expressions in Japanese to English MT and its effects. In *Proceedings of the 2nd International Workshop on Controlled Language Applications (CLAW)*, pages 62–75, Pennsylvania, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, Massachusetts, USA.

Yanli Sun, Sharon O'Brien, Minako O'Hagan, and Fred Hollowood. 2010. A novel statistical pre-processing model for rule-based machine translation system. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT)*, Saint-Raphaël, France.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais*. Didier, Paris, trans. and ed. by J. C. Sager & M.-J. Hamel (1995) as *Comparative Stylistics of French and English: A Methodology for Translation*. John Benjamins, Amsterdam.

Sanja Štajner and Maja Popović. 2018. Improving machine translation of English relative clauses with automatic text simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 39–48, Tilburg, Netherlands.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 508–514, Geneva, Switzerland.

Takehiko Yoshimi. 2001. Improvement of translation quality of English newspaper headlines by automatic pre-editing. *Machine Translation*, 16(4):233–250.

Zhongyuan Zhu. 2015. Evaluating neural machine translation in English-Japanese task. In *Proceedings of the 2nd Workshop on Asian Translation (WAT)*, pages 61–68, Kyoto, Japan.

Explicitation strategy	Total	Example of ST pre-editing	MT output
Information addition	142	12日は台湾の休日のため休場。 *12-nichi wa taiwan no kyujitsu no tame kyujo.*	The twelfth is a holiday in Taiwan.
		→ 12日は台湾の休日のため**株式市場は**休場。 *12-nichi wa taiwan no kyujitsu no tame **kabushiki shijo wa** kyujo.*	**The stock market** was closed on the twelfth due to a holiday in Taiwan.
Use of clear relation	103	来院しなくても10日前後で登録のクレジットカードから引き落としを行います。 *Raiin-shinakutemo toka zengo de touroku no kurejitto kado kara **hikiotoshi o okon-aimasu.***	Withdraw from your registered credit card in about 10 days without visiting the hospital.
		→ 来院しなくても10日前後で登録のクレジットカードから引き落としが行われます。 *Raiin-shinakutemo toka zengo de touroku no kurejitto kado kara **hikiotoshi ga okon-awaremasu.***	Even if you do not visit the hospital, your credit card will be debited in about 10 days.
Use of narrower sense	54	採尿と採便を出してください。 *Sai-nyo to sai-ben o **dashite** kudasai.*	Please **collect** urine and feces.
		→ 採尿と採便を提出してください。 *Sai-nyo to sai-ben o **teishutsushite** kudasai.*	Please **submit** urine and stool samples.
Normalisation	30	単位は億円。 *Tan'i wa oku en.*	**Figures are in billions of yen**.
		→ 単位は億円です。 *Tan'i wa oku en **desu**.*	**The unit is 100 million yen**.

Table 7: The number of instances and an example of each explicitation strategy for pre-editing ST with MT outputs.

A   Details of Explicitation Strategy

Table 7 shows the statistics and examples of each subcategory of the explicitation strategy. A total of 329 pre-editing instances of the explicitation strategy can be further classified into four subcategories: information addition, use of clear relation, use of narrower sense, and normalisation.

The example of the information addition illustrates the insertion of a subject '*kabushiki shijo wa*' ('stock market'), which is implicit in the preceding ST. The example of the use of clear relation shows that the relation between the subject and object can be clarified by using the nominative case marker '*ga*' instead of the accusative one '*o*' and accordingly changing the voice of the main clause. As a result, the inappropriate imperative construction 'Withdraw from ...' in the MT output is changed to the correct passive construction 'will be debited.' In the example of the use of narrower sense, the verb '*dashite*,' which has multiple meanings such as 'put,' 'take,' and 'send,' was replaced with the verb '*teishutsushite*,' which has a narrower range of meaning and was correctly translated as 'submit.' In the example of normalisation, the elliptic sentence-ending was completed with a normal structure '... desu.' This operation led to not only

the improvement of the sentence construction, but also the semantic correctness in the MT output ('billions'→'100 million').