Evaluating the suitability of human-oriented text simplification for machine translation

Rei Miyata
Nagoya University
miyata@nuee.nagoya-u.ac.jp

Midori Tatsumi
Rikkyo University
midori.tatsumi@rikkyo.ac.jp

Abstract

We present the results of an experiment to evaluate the suitability of simplified text as a source for machine translation (MT). Focusing on Japanese as the source language, we first proposed a simplest possible rule set to write text that can be easily understood by language learners and children. Following this rule set, we manually rewrote expository sentences concerning Japanese cultural assets in simplified Japanese, through two steps: (1) splitting long sentences into short complete sentences, and (2) further simplifying these. We then conducted a human evaluation to assess the quality of the English MT outputs of the original, split, and simplified sentences. The results indicated the potential of simplified text as an effective source for MT, demonstrating that nearly 80% of the raw MT outputs achieved usable quality. The qualitative analyses also revealed occasional side effects of simplification and fundamental difficulties for MT.

1 Introduction

Text simplification is the process of reducing the complexity of the sentence structure and difficulties of the words in a given text. The applications of text simplification vary from reading aids for human readers to preprocessing for natural language components, such as machine translation (MT). While automatic text simplification techniques have been proposed, with the effectiveness demonstrated on certain evaluation tasks, many practical attempts, such as Simple English Wikipedia, rely mostly on manual text simplification with some writing guidelines. In this context, we have developed a simplified Japanese rule set for non-professional writers, which requires the rules to be simple for such writers to follow. Our rule set is intended for writing expository text on Japanese cultural assets. This is challenging, as such texts contain many culture-specific technical terms that are difficult to simplify, even for human writers.

Although the primary purpose of our simplified Japanese is to enhance the text readability for those with limited Japanese proficiency, such as language learners and children, we are also interested in investigating the machine translatability of a simplified source text, especially considering the recent developments of neural MT (NMT) technology. To date, little effort has been invested in examining the compatibility between text simplification approaches for human readers and MT in detail. Here, three major questions arise:

1. To what extent can manual text simplification improve MT outputs?
2. What types of simplification operations are effective or ineffective for improving the MT quality?
3. What types of translation difficulties remain even after the source text is simplified?

Therefore, in this study, focusing on Japanese and English as the source and target languages, respectively, we address these questions by proposing simplified Japanese for human readability and evaluating the suitability of the simplified text as a source for MT. To investigate the effect of the simplification process in detail, we decompose it into two op-

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 147-155, Hakodate, Japan, September 13-15, 2019

erations: (1) splitting long sentences into short complete sentences, and (2) further simplifying these. To test the suitability of this simplification for MT, we evaluate the MT output quality and diagnose the MT errors.

We discuss related work in Section 2, and introduce our guidelines for simplifying Japanese in Section 3. Section 4 describes the process and product of the manual simplification of text. We explain our experimental setup in Section 5, and present our results with in-depth analyses in Section 6. Finally, Section 7 concludes the paper and proposes future research directions.

2 Related work

Automatic text simplification has been tackled in the natural language processing research field for various purposes and languages (Siddharthan, 2014; Shardlow, 2014). However, full automation remains difficult, particularly for human-oriented text simplification tasks, which require the produced text to be of high quality. In many practical applications, human writers conduct text simplification tasks by means of authoring guidelines and technological aids. For instance, Wikipedia provides guidelines and introduces several existing support tools for writing simplified English versions of regular Wikipedia pages.[1] The guidelines specify vocabulary lists such as Basic English 850/1500 and simple sentence structures. They also define the preferred use of voice (active voice) and tenses (past, present or future only).

One of the most widely-acknowledged simplified Japanese rule sets is *Yasashii Nihongo*, or 'Easy Japanese', proposed by the Sociolinguistics Laboratory at Hirosaki University.[2] This consists of 12 writing rules, which restrict the vocabulary and regulate certain types of complex structures, such as long sentences and double negation. Because the original purpose of Easy Japanese was to provide foreign residents in Japan with emergency information, the vocabulary restrictions are rather strict, with about 1400 basic words, which corresponds to the Japanese-Language Proficiency Test (JLPT) Grade 3 and Grade 4.[3]

Inspired by this rule set, several human-oriented simplified Japanese guidelines have been developed, such as those for disseminating local community information (Iori, 2016) and writing news report scripts (Tanaka et al., 2013). While the details of these simplified languages differ depending on the purpose and audience, the shared core idea is to prescribe a vocabulary list and restrict complex sentence structures, which basically corresponds to two major subtasks of (automatic) text simplification: lexical and syntactic simplification (Shardlow, 2014; Saggion, 2017).

One of the most important aspects of a practical implementation of simplification lies in the simplicity of the guidelines. Some simplified languages that are mainly utilised by professional writers specify detailed usages of lexicons, grammars and styles. For example, ASD-STE100 (ASD, 2017), also recognised as a controlled language, defines 53 writing rules and a dictionary of approved and not-approved words for writing technical documentation. However, for non-professional writers, the guidelines themselves should be sufficiently simple for utilisation.

MT-oriented text simplification has also been undertaken (Hung et al., 2012; Štajner and Popovic, 2016; Štajner and Popovic, 2018). For example, Štajner and Popovic (2016) employed two automatic text simplification systems to produce lexically and syntactically simplified versions of source text for English-to-Serbian statistical MT, and evaluated the MT outputs in terms of the fluency, adequacy and post-editing effort. While these studies demonstrate the efficacy of automatic text simplification techniques for MT applications, two major issues remain: (1) human readability is not explicitly taken into account, and (2) the potential gain in MT quality when manual text simplification is fully performed is not measured.

In the research field of controlled language, several evaluation experiments have examined the compatibility or commonality between human-oriented

[1] Simple English Wikipedia, https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

[2] http://human.cc.hirosaki-u.ac.jp/kokugo/EJ9tsukurikata.ujie.htm

[3] JLPT Grade 3 and Grade 4 correspond to current versions of N4 (the ability to understand basic Japanese) and N5 (the ability to understand some basic Japanese). https://www.jlpt.jp/e/about/levelsummary.html

and MT-oriented controlled language rules (O'Brien and Roturier, 2007; Aikawa et al., 2007; Hartley et al., 2012; Miyata et al., 2015). However, these studies tend to focus on structural and stylistic aspects of technical documents. The effect of vocabulary restriction, which is a major task of text simplification, has not been significantly investigated. Moreover, NMT systems has not yet been examined.

As Koehn and Knowles (2017) demonstrated, despite its recent advancements NMT still faces difficulties in dealing with low-frequency words and long sentences, among others. This naturally motivates us to assume that text simplification that restricts vocabulary and sentence complexity can be helpful to enhance MT quality, even if it is intended for human readability. However, as noted by Hartley et al. (2012) and Miyata et al. (2015), there are incompatibilities between human readability and machine translatability. Therefore, an in-depth analysis of the suitability of human-oriented text simplification for MT is required to understand its potential and limitations.

3 Simplified Japanese rule set

There is no single standard rule set for simplified Japanese. Variations exist to adjust the level of Japanese depending on the type of information to be conveyed and the target audience, as mentioned in Section 2. For writing about cultural assets, at least an upper-intermediate vocabulary level will be required. On the other hand, the sentence structure could be limited to a basic level.

In general, simplified Japanese is written by Japanese language teachers, or those who are trained to author in simplified Japanese. However, our aim is to create a rule set that is sufficiently simple to be understood and followed by lay people, namely those that are neither professional linguists nor Japanese language teachers. Therefore, we avoid using grammatical terms or complicated linguistic concepts when setting the rules. Essentially, our rule set consists of just the following three rules.

Rule 1: Present no more than one idea per sentence.

Rule 2: Specify the subject as far as possible, and if the subject is implied then use the passive tense.

Rule 3: Use only the vocabulary and Kanji (Chinese characters) of up to JLPT Grade 2.

JLPT no longer has an official list of vocabulary and Kanji for each level. Thus, we employed the equivalent list available on the website of the Faculty of Humanities and Social Sciences at Hirosaki University.[4] This list contains 3,708 words for Grade 2, 688 words for Grade 3 and 740 words for Grade 4, where smaller grades indicate a higher level.

It should be noted that there are cases in which we could not rewrite a sentence to strictly conform to these rules. For example, there are sentences that are left without a subject, as specifying a subject for every predicate can make some Japanese sentences sound unnatural. We also left proper nouns as they are, even if they are not found in the list of vocabulary and Kanji up to Grade 2.

4 Simplification

4.1 Dataset

We collected 1,274 Japanese sentences from leaflets on historical buildings and houses that have officially been designated as Japanese cultural assets. These leaflets are available either as printed matters at the physical sites, or in downloadable electronic format (PDF) on their official websites.

In the collected text, we identified the following nine topics: *Style and features*, *History and episodes*, *Owner and resident*, *Architect*, *Environment*, *Artefacts and objects*, *Access information*, *Captions and titles*, and *Other*. We categorised each sentence according to the topics, because the topic is an important determining factor for the grammatical construction of a sentence. For example, many of the sentences in the *Style and features* category are descriptive, and can be written in the form of 'X is/are ...' and 'There is/are ...', while the majority of the sentences in the *History and episodes* category are anecdotal and expressed in past tense. For the present study, as a starting point we focus on *Style and features*, which is the most dominant topic in the collected data.

Some of the sentences were comprised of a mixture of different categories. We eliminated such cases, and obtained 206 sentences for the original Japanese source text (**ST-org**).

[4] http://human.cc.hirosaki-u.ac.jp/kokugo/CATtwo/youziyougoziten/youziyougoziten_96_165.pdf

4.2 Simplification process

Rewriting was performed by one of the authors of this paper, who is not a teacher of Japanese language nor trained in writing simplified Japanese. This is preferable, as our presumed writers do not necessarily have such qualifications. ST-org were rewritten according to our simplified Japanese rules in two steps: sentence splitting and further simplification.[5] In principle, sentence splitting covers Rules 1 and 2 and further simplification covers Rules 2 and 3.

Sentence splitting To fulfil Rule 1 of our simplified Japanese, we split the original sentences as required, such that each sentence presents only one idea. For example, 敷鴨居などには白檀が使われ、暖房時には芳香が漂いました。('Sandal wood was used for "Shikikamoi", and it smelled good when heating the room.") was split into two shorter sentences at the location of 'and'. Some splitting operations required the supplementation of linguistic elements, such as subjects and objects, to follow Rule 2. In addition, we tried to utilise the simplest sentence patterns as far as possible. For example, complex predicates such as ～が設置されています ('... is installed') and ～が施されています ('... is in place') have been changed to ～があります ('there is ...'). For the 206 ST-org sentences, we obtained 509 corresponding split sentences (**ST-split**).

Further simplification Based on Rule 2, we further specified a subject for each predicate, and when this was not possible we changed the active voice to the passive voice. For example, 床の間には掛け軸を飾っていました。 has no subject, and a literal translation would be 'Used to display a painting in the alcove.' This was changed to 床の間には掛け軸（絵）が飾られていました。, meaning 'A painting used to be displayed in the alcove.' At this stage, according to Rule 3, we changed the words and expressions such that the sentences consisted as far as possible of only vocabulary and Kanji up to JLPT Grade 2. For example, we changed 採光性に優れています (literally meaning 'excellent in daylighting') to 光がたくさん入ります ('a lot of light enters'). We call this final version of source text **ST-simple**, which consists of 511 sentences. The reason that the number of sentences in ST-simple is slightly larger than that in ST-split is that in rare cases there was a need to further split sentences to simplify them.

[5]Recent studies on building Japanese simplification resources, such as Maruyama and Yamamoto (2018), tend to focus on lexical simplification, a subset of the whole process.

	ST-org		ST-split		ST-simple	
	#	%	#	%	#	%
OOV	1064	21.62	1120	18.76	453	7.58
Grade 2	712	14.47	867	14.52	1220	20.41
Grade 3	371	7.54	529	8.86	620	10.37
Grade 4	1500	30.48	1999	33.48	2110	35.31
F/S	1275	25.90	1456	24.38	1573	26.32
Total	4922	100	5971	100	5976	100

Table 1: Statistics of vocabulary level (OOV > Grade 2 > Grade 3 > Grade 4 > F/S: Functional words/Symbols)

4.3 Vocabulary level of simplified Japanese

Table 1 presents the statistics for the vocabulary levels of words in each of the source versions (ST-org, ST-split and ST-simple). The number of total words increased from ST-org to ST-split, because we supplemented necessary words and did not omit information as far as possible when splitting sentences.

Out-of-vocabulary (OOV) can be regarded as difficult words above the Grade 2 level of JLPT. The ratio of OOV was reduced considerably from ST-split (18.76%) to ST-simple (7.58%), which demonstrates the effect of lexical simplification, although it was not possible to completely eliminate OOV even after manual simplification.

We also observe that the ratio of Grade 2 words considerably increased from ST-split (14.52%) to ST-simple (20.41%). This means that most of the OOV were changed to Grade 2.

5 Experimental setup

We translated the three versions of the Japanese source text using Google Translate,[6] to obtain three versions of English target text: **MT-org**, **MT-split** and **MT-simple**. The resulting English translations were then evaluated by a professional linguist, whose native language is Japanese and who has 10 years of experience in professional Japanese to English translation. The reason we chose a native Japanese speaker was that the Japanese original source sentences are loaded with culture-specific terms that need to be understood without facing a cultural barrier. Furthermore, it was not necessary or desirable to review the translation in terms of

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

Good	The information of the source text has been completely translated and there are no grammatical errors in the translation. There may be some unnatural word choices and/or phrasings, but these would not hinder understanding of the meaning.
Fair	There are some minor errors in the translations of less significant parts of the source text, but the meaning of the source text can easily be understood.
Acceptable	Some of the source text is omitted or incorrectly translated, but the core meaning can still be understood with some effort.
Incorrect	Even the core meaning of the source text is not conveyed.
ST unclear	It is impossible to assess the quality of the MT output because of incomprehensible/ambiguous words and/or expressions in the source text.

Table 2: Evaluation criteria

	MT-org		MT-split		MT-simple	
	#	%	#	%	#	%
Good	53	25.73	240	47.15	317	62.04
Fair	8	3.88	26	5.11	37	7.24
Acceptable	17	8.25	35	6.88	49	9.59
Incorrect	65	31.55	114	22.40	76	14.87
ST unclear	63	30.58	94	18.47	32	6.26
Total	206	100	509	100	511	100

Table 3: MT quality

the naturalness or stylistic appropriateness from the viewpoint of a native English speaker.

The 1,226 sentences comprising the three versions were put in a random order to prevent the evaluator from deducing their meanings from the surrounding sentences. We asked the evaluator to rate the quality of the English translations using the five grades shown in Table 2, which are versions of the acceptability evaluation grades used by Goto et al. (2013) modified for the purpose of the present study.

The grade *ST unclear* was added to isolate cases in which the source text contains highly technical terms that lay people, even adult native Japanese speakers, would not understand. In such cases, we may not be able to expect a meaningful evaluation.

The evaluator was also asked to highlight sections in the source and target texts that were incomprehensible, enabling us to qualitatively diagnose the translation difficulties.

6 Results and analyses

6.1 Overall results for MT quality

Table 3 summarises the results of the quality evaluation of the English translation. Approximately 30% of the MT-org sentences are rated as *ST unclear*. The majority of the elements in Japanese source sentences reported as incomprehensible are technical terms relating to architecture or Japanese culture (technical terms related to a tea ceremony, for example). After splitting the sentences, the proportion of *ST unclear* is reduced to less than 20% in MT-split. This is because one or some of the split sentences still contain the same terms, while others become free of them. For MT-simple, only 6.26% are rated as *ST unclear*, because most of the technical terms have been replaced with simpler words or explanatory expressions using simple words.

Simply splitting a sentence to allow each sentence to contain only one idea can double the rate of producing a *Good* translation (25.73% to 47.15%), and employing simple words and expressions can further increase the ratio to 62.04%. Similarly, the percentage for *Incorrect* decreases from 31.55% to 22.40% by splitting the sentences. Further simplification can reduce the percentage of *Incorrect* to 14.87%.

We consider translations with the grades *Good*, *Fair* and *Acceptable* as 'usable', as at least the core meaning of the source text is conveyed. This means that while less than 40% of the ST-org sentences can produce usable translations, approximately 60% of those in ST-split and almost 80% of those in ST-simple can. This result illustrates the high suitability of human-oriented text simplification for MT.

6.2 Analysis of simplification operations

6.2.1 Sentence splitting

Among the 63 MT-org sentences rated as *ST unclear*, there were no cases in which all corresponding MT-split sentences obtained *Good* or *Fair* ratings. This is expected, because as mentioned in Section 6.1 the reasons for incomprehensibility mostly relate to technical terms, which remain even after splitting a sentence.

There are 65 cases in which MT-org sentences received *Incorrect* ratings, and in 12 cases all corre-

ST-org	十字の形でその四方に窓があり日差しを多く取り込めるデザインになっています。
MT-org	It has a window in the form of a cross and has a design that can capture a lot of sunlight.
ST-split	この部屋は十字の形です。/ この部屋の四方には窓があります。/ 日差しを多く取り込めるデザインになっています。
MT-split	**This room** is in the form of a cross. / There are windows on all sides of **this room**. / It is designed to capture a lot of sunshine.

Table 4: Example of the positive effect of sentence splitting

ST-org	天井板には美しい木目を活かした木板が組み合わされています。
MT-org	The ceiling board is a combination of **wood boards that take advantage of beautiful wood grain**.
ST-split	天井板には木板が組み合わされています。/ 美しい木目が活かされています。
MT-split	**A wood board** is combined with the ceiling board. / **Beautiful wood is used**.

Table 5: Example of the negative effect of sentence splitting

ST-org	1 階居間の暖炉には、アールヌーボー風のタイルを使い、ケヤキ材の前飾りがついている。
MT-org	**The fireplace** in the living room on the first floor is made of Art Nouveau-style tiles and decorated with a zelkova wood front decoration.
ST-simple	1 階の居間の暖房には、アールヌーボーのタイルが使われています。/ 飾りは「ケヤキ」という木でできています。
MT-simple	Art Nouveau tiles are used to **heat** the living room on the first floor. / The decoration is made of a tree called “keyaki”.

Table 6: Example of the negative effect of lexical simplification

sponding MT-split sentences obtained *Good* or *Fair* ratings. The main reason for this is that the ill-formedness of sentences is corrected by splitting them into shorter ones. Table 4 presents an example; the complex dependency relations were resolved, and the missing subject この部屋 (‘this room’) was supplemented, as a result of applying Rules 1 and 2, respectively, in the sentence splitting step.

However, there are cases in which splitting the source sentence degrades the quality of the MT outputs. Table 5 presents an example. Here, the sentence was split to prevent the noun 木板 (‘wood boards’) from having the long adjective clause 美しい木目を生かした (‘that take advantage of beautiful wood grain’), which was actually translated correctly in MT-org. In this example, it appears that separating the latter part caused a mistranslation of the relationship between the ‘ceiling board’ and ‘wood boards’. Excessive splitting of a sentence may reduce contextual information within the sentence, leading to the degradation of the MT output.

6.2.2 Further simplification

Among the 208 *Incorrect*/*ST unclear* cases in MT-split, 132 became *Good*/*Fair*/*Acceptable* in MT-simple. The reasons for the majority of the improvements in the MT outputs lie in the rephrasing of technical terms using their hypernyms or explanatory expressions. For example, 袖塀, the name of a special type of wall, has been replaced with 壁, which simply means ‘wall’. In addition, 板透し彫, the name of a special type of decoration, has been replaced with the explanatory expression 木で作った模様, meaning ‘decorations made of wood’. This shows that Rule 3 (Use only the vocabulary and Kanji of up to JLPT Grade 2) is not only beneficial for human readers, but also for MT.

However, there are 32 cases in which further simplification degraded grades from *Good*/*Fair*/*Acceptable* to *Incorrect*/*ST unclear*. Table 6 presents an example of the harmful effect of replacing the term 暖炉 (‘fireplace’) with the presumably simpler term 暖房 (‘heating’). This mistranslation was caused by the equivocality of the word 暖房, which can mean both ‘heating equipment’ and ‘the act of heating’.

Current MT systems have significantly larger vocabularies than those used in human-oriented text simplification. In other words, most general words, even if they are difficult, can be covered by MT systems. In summary, the simplification of rare technical terms is effective for both human and MT applications, but simplifying general words may result in ambiguous words, having an adverse effect on MT.

6.3 Analysis of factors for MT quality

6.3.1 Relation between source sentence characteristics and MT quality

Our motivation for investigating the suitability of text simplification for MT is based on the assumption that long sentences and difficult words can be major factors in degradation in MT quality. Here, we further explore the relation between source sentence characteristics and MT quality.

Table 7 presents correlation scores (Spearman's ρ and Kendall's τ), demonstrating the weak correlations between the MT quality and the numbers of words, characters and OOV in a sentence. The number of OOV is a slightly better indicator than the sentence length for estimating the MT quality.

Figures 1 and 2 present box plots for the sentence length and number of OOV for each MT quality grade. The bold vertical line in each box indicates the median. The majority of *Good*/*Fair*/*Acceptable* MT outputs are produced from source sentences that are no more than 15 words in length and contain no more than two OOV words.

However, some rather long sentences resulted in *Good* quality translations. Table 8 presents an example. In ST-org, the subject この建物 ('the building') only appears once, while there are two predicates 'is ...'. While Japanese sentences often omit the subject, and even change the subject in the middle of a sentence without clearly indicating this change, in this example the subject この建物 is present at the beginning, and is the subject for both predicates. The MT system successfully supplements 'it' to continue the sentence, although it failed to add 'and' before the pronoun. These examples indicate that the source sentences do not necessarily have to be short, so long as they employ grammatically correct subject–predicate combinations.

6.3.2 Remaining difficulties for MT

Finally, we focus on the 76 cases in which MT-simple sentences received *Incorrect* ratings. Referring to the highlighted sections of text that were judged as incomprehensible by the evaluator (see Section 5), we identified a total of 87 critical MT errors, ignoring minor grammatical, orthographic and stylistic errors that do not impair the core meaning of the source text. Based on the MT error taxonomies presented in Costa et al. (2015) and Popovic (2018), we classified the errors as shown in Table 9.

	Spearman's ρ	Kendall's τ
# of words	0.278	0.217
# of characters	0.220	0.170
# of OOV	0.328	0.271

Table 7: Correlations with MT quality

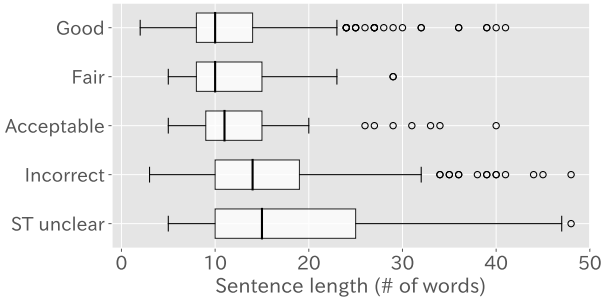

Figure 1: Relation between MT quality and # of words

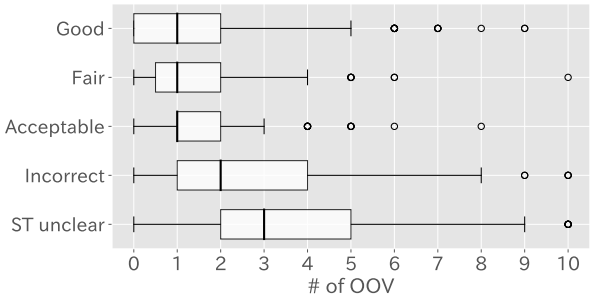

Figure 2: Relation between MT quality and # of OOV

The most frequent error type is the mistranslation of technical terms, including proper nouns. By nature, it is difficult for NMT to correctly handle rare words (Li et al., 2016; Koehn and Knowles, 2017). Although we reduced the technical terms as far as possible through the simplification process, it was impossible to write text on cultural assets without any technical terms. Nevertheless, we can predict possible MT errors if we are aware of the existence of such words, which enables the strategic deployment of post-editing.

The second most frequent error type is the confusion of senses. For example, in many cases 木 was translated as 'tree', although the correct translation was 'wood'. Human translators can easily disambiguate the senses using subtle clues in the text and common knowledge. As detailed contextual information tends to be avoided in simplified text, word

ST-org	この**建物**は、木造総 2 階建て住宅で、細部にはカーペンターゴシック様式の意匠が見られる、19 世紀後半のアメリカ郊外住宅の特色を写した質素な外国人住宅です。
MT-org	**The building** is a two-story wooden house with a carpenter gothic design in every detail, **it** is a frugal foreign house that features the characteristics of an American suburb in the late 19th century.

Table 8: A long ST-org that produced a *Good* MT output

Level 1	Level 2	Level 3	#
Lexis	Mistranslation	Common words	5
		Technical terms	39
	Omission		3
	Untranslated		1
Semantic	Confusion of senses		26
	Mistranslation	Subjects	4
		Others	9

Table 9: Classification of remaining MT errors

sense disambiguation remains a major issue for MT. One solution is domain-adaptation. In a general domain, 'tree' is the most probable translation, while in this particular domain of cultural assets, 'wood' would be the most probable. Thus, retraining MT using in-domain data would be effective if sufficient data is available. Another solution is the use of concrete words. For example, 木材 is likely to be translated as 'wood', as this word has a smaller range of meaning than 木. Although 木材 is more difficult for the target audience than 木, it is still in the vocabulary list for our simplified Japanese.

Although not frequent, the mistranslation (or misidentification) of subjects is noteworthy. For example, 山小屋のような感じがします。 is translated as 'I feel like a mountain hut.' The correct translation is 'It feels like a mountain hut.' In this case, the lack of a subject caused the insertion of the incorrect subject 'I' by the MT system. Although it is possible for human writers to supplement a subject such as これ ('it') or このデザイン ('this design') in the source, repeated use of the same subject may be regarded as unnatural in Japanese. To cope with the incompatibility between source naturalness and machine translatability, we need to incorporate an additional process to further modify the human-oriented simplified source text such that it can contains the necessary subjects to produce a better MT result.

7 Conclusion

In this study, we have proposed a simple rule set for simplified Japanese for human readability, and examined the suitability of simplified text as a source for machine translation (MT). Focusing on expository sentences on Japanese cultural assets, we manually conducted a simplification task in two steps: (1) splitting long sentences into short complete sentences, and (2) further simplifying them. The Japanese-to-English neural MT outputs of the original, split and simplified sentences were manually evaluated in terms of the MT quality.

The experimental results demonstrated the strong potential of human-oriented text simplification for MT purposes, showing that almost 80% of the raw MT outputs achieved a usable quality, among which approximately 80% were of *Good* quality, i.e., the information of the source text was completely translated without grammatical errors. Although the fact that structural and lexical simplification helps to improve the MT quality is not surprising per se, this result reveals the detailed gains we can expect to obtain from simplification.

We also conducted in-depth analyses of the results. The findings can be summarised as follows:

- Splitting sentences is effective when this can resolve ill-formed structures, while excessive splitting may degrade the MT outputs.
- Avoiding rare technical terms is generally effective, while lexical simplification sometimes makes the source text simple but ambiguous.
- Technical terms, word sense ambiguity and a lack of subjects are critical difficulties for MT, which remain even after the text is simplified.

In future work, we intend to tackle the identified difficulties, specifically technical terms and lacking subjects. For technical terms, we plan to develop a tool to generate alternative expressions, such as hypernyms and explanatory phrases. For lacking subjects, we will introduce a semi-automatic process to add subjects necessary only for MT.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP17K00466 and JP19K20628.

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.

ASD. 2017. ASD Simplified Technical English. Specification ASD-STE100, Issue 7. `http://www.asd-ste100.org`.

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Isao Goto, 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 NTCIR Conference*, pages 260–286, Tokyo, Japan.

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.

Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. 2012. Sentence splitting for Vietnamese-English machine translation. In *Proceedings of the 4th International Conference on Knowledge and Systems Engineering (KSE)*, pages 156–160, Danang, Vietnam.

Isao Iori. 2016. The enterprise of Yasashii Nihongo: For a sustainable multicultural society in Japan. *Jinbun-Shizen Kenkyu*, (10):4–19.

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.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2852–2858, New York, USA.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified corpus with core vocabulary. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 1153–1160, Miyazaki, Japan.

Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. 2015. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proceedings of the Machine Translation Summit XV*, pages 90–103, Miami, Florida, USA.

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.

Maja Popovic. 2018. Error classification and analysis for machine translation quality assessment. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*. Springer.

Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications: Special Issue on Natural Language Processing*, 4(1):58–70.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics: Recent Advances in Automatic Readability Assessment and Text Simplification*, 165(2):259–298.

Hideki Tanaka, Hideya Mino, Shinji Ochi, and Motoya Shibata. 2013. News services in simplified Japanese and its production support systems. In *Proceedings of the International Broadcasting Convention 2013 (IBC)*, Amsterdam, The Netherlands.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? *Baltic Journal of Modern Computing*, 4(2):230–242.

Sanja Štajner and Maja Popovic. 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, The Netherlands.