Attainable Text-to-Text Machine Translation vs. Translation: Issues Beyond Linguistic Processing

Atsushi Fujita atsushi.fujita@nict.go.jp
National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan

Abstract

Existing approaches for machine translation (MT) mostly translate a given text in the source language into the target language, without explicitly referring to information indispensable for producing a proper translation. This includes not only information in the textual elements and non-textual modalities in the same document, but also extra-document and non-linguistic information, such as norms and skopos. To design better translation production workflows, we need to distinguish translation issues that could be resolved by the existing text-to-text approaches from those beyond them. To this end, we conducted an analytic assessment of MT outputs, taking an English-to-Japanese news translation task as a case study. First, examples of translation issues and their revisions were collected by a two-stage post-edit (PE) method: performing a minimal PE to obtain a translation attainable based on the given textual information and further performing a full PE to obtain an acceptable translation referring to any necessary information. The collected revision examples were then manually analyzed. We revealed the dominant issues and information indispensable for resolving them, such as fine-grained style specifications, terminology, domain-specific knowledge, and reference documents, delineating a clear distinction between translation and the translation that text-to-text MT can ultimately attain.

1 Introduction

Translation is not a purely linguistic process (Vermeer, 1992) but also the process of producing a document in the target language that plays the same role (has the same effect) as the given source document written in the source language. When translating a given document, translators refer not only to the textual elements in the document, but also to the role of each textual element (e.g., running text, section title, table element, and caption), other non-linguistic elements (e.g., figures and formulae), and their structure. To produce a translation, we also need some extra-document and non-linguistic information, such as the *norms* specific to the register of the document and corresponding target sub-language (Toury, 1978), the objective and the intended usages of translation, i.e., *skopos* (Vermeer, 2004), and various specifications (Melby, 2012) designated by the translation client if any.

Despite the requirements a (proper) translation must satisfy, techniques for machine translation (MT) have been developed by regarding the task of translation as *text-to-text transfer*. Until very recently, most studies have performed a text-to-text MT for each text *segment*,[1] even though a sequence of perfect segment-level text-to-text translations does not necessarily qualify as a proper translation. Recent studies on neural MT (NMT) have addressed issues beyond

[1] In this paper, we use "segment" for the unit of inputs for MT systems rather than "sentence," because a segment is not necessarily composed of a single sentence, but can often be multiple sentences or non-sentential textual fragments.

this formulation, exploiting further information such as document-level textual context (Voita et al., 2018, 2019; Lopes et al., 2020) and other modalities (Barrault et al., 2018). There are also several focused studies on exploiting extra-document and non-linguistic information. However, such information has not been extensively discussed. As a result, in translation production workflows at translation service providers (TSPs), where MT outputs are treated as draft translations, heavy human labor is necessary to fill the gap between MT outputs and translations in addition to resolving issues at the text-to-text level, for instance, by manual post-editing (PE).

To design and establish more practical ways of exploiting MT systems in translation production workflows as well as to discuss how to make MT systems more useful, we need to understand what lies in the gap between a translation that text-to-text processing can attain and a truly acceptable translation. Moreover, this should be shared among not only translators but also MT researchers and MT users. From this point of view, this paper presents our analytic assessment of MT outputs, taking an English-to-Japanese news translation task as a case study. First, we obtained segment-level text-to-text translation by resolving translation issues in MT outputs. At this stage, a minimal PE was performed referring only to each source segment isolated from any other information, and thus the results represent what segment-level text-to-text MT systems can ultimately attain. Then, the document-level full PE (ISO/TC37, 2017) in the succeeding stage resolved all the remaining issues, i.e., those issues lying in the gap between acceptable segment-level text-to-text translation and proper translation. Finally, the collected revision examples were manually analyzed based on an issue classification scheme. This revealed several dominant issues as well as the information indispensable for resolving them.

The remainder of this paper is organized as follows. Section 2 summarizes related work in translation studies and MT. Section 3 presents the material for our case study. Section 4 describes our workflow, designed for collecting translation issues that cannot be solved by text-to-text processing. Section 5 presents our analytic assessment of translation issues, which relies on an existing issue typology, and explains the dominant issues as well as several types of extra-document and/or non-linguistic information that must be used to solve them. Section 6 describes future research directions and advice for non-expert MT users, and Section 7 concludes the paper.

2 Related Work

In the literature of translation studies, linguistic approaches to translation have been criticized (Kenny, 2001), and the *equivalence* of a source document and a target document has been studied from a diverse range of aspects. In a seminal work, Nida (1964) claimed the necessity of equivalence of recipients' reactions when reading source and target documents. Chesterman (1997) compiled a typology of translation strategies adopted to guarantee the equivalence when producing a translation. His syntactic and semantic strategies can be explained (and potentially realized) referring only to textual information in the source document and linguistic knowledge in general. In contrast, some of his pragmatic strategies, such as cultural filtering and illocutionary changes, require extra-document and/or non-linguistic information.

Some of the kinds of information that must be referred to for producing a proper translation, including terminologies and style specifications, are mentioned in the translation workflow standard, ISO 17100 (ISO/TC37, 2015). Other items are mentioned in existing criteria for quality assurance, such as the Multidimensional Quality Metrics (MQM)[2] and the Dynamic Quality Framework (DQF).[3] Reference sources, such as translation memories and bilingual concordancers, and other access to past translations are valuable assets for improving efficiency in personal practices and workflows in TSPs. However, there is neither a comprehensive inven-

[2] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

[3] https://www.taus.net/data-for-ai/dqf

tory of references, nor a common view of the extent of the necessity and availability of each reference depending on the given skopos.

Recent advances in MT go beyond segment-level and/or text-to-text processing. For instance, Voita et al. (2019) focused on several discourse-level issues, i.e., deixis, lexical cohesion, and ellipsis, occurring in segment-level text-to-text MT. Following studies proved that context-aware decoding that refers to several preceding segments better handles these linguistic phenomena (Lopes et al., 2020). There are several focused studies on exploiting extra-document and non-linguistic information, including terminologies (Arthur et al., 2016; Hasler et al., 2018), politeness (Sennrich et al., 2016a), domain (Chu et al., 2017; Kobus et al., 2017; Bapna and Firat, 2019), style (Niu et al., 2017; Michel and Neubig, 2018b), markups (Chatterjee et al., 2017; Hashimoto et al., 2019), and external lexical knowledge (Moussallem et al., 2019). However, the information indispensable for producing a proper translation have not been thoroughly studied. More importantly, no work guarantees to perfectly reflect such information.

The MT community has benefited from manual analyses of translation issues[4] caused by MT systems. Existing methodologies for analyzing translation issues in MT outputs can be two-fold: (a) comparisons of independent products, i.e., MT outputs and human translations (Popović and Ney, 2011; Irvine et al., 2013; Toral, 2020), and (b) annotations of the issues in MT outputs according to pre-determined issue typologies, such as MQM and DQF (Lommel et al., 2015; Ye and Toral, 2020; Freitag et al., 2021). The issues identified in the former approach contain both true errors and preferential differences, i.e., alternative acceptable translations independently selected by MT systems and humans. The latter approach enables us to clearly separate them. For instance, past studies (Hardmeier, 2014; Scarton et al., 2015; Voita et al., 2019) analyzed outputs of segment-level text-to-text MT, showed the limitation of that approach, and encouraged the research on document-level MT. However, they discussed only the differences between two text-to-text approaches. Issues beyond the text-to-text processing, such as those related to extra-document and/or non-linguistic information, have seldom been mentioned (Castilho et al., 2020), and no focused and empirical analysis has been conducted.

3 Subject of Our Case Study

Our focus in this paper is to clarify the types of extra-document and/or non-linguistic information that are indispensable for producing a translation. Among several translation tasks, this paper takes an English-to-Japanese news translation task as a case study and presents our in-depth analysis. We chose it for two reasons. First, despite the high demand for it, the task is still very difficult, since the two languages are linguistically distant and used in substantially different cultures (cf. English-to-German studied by Scarton et al. (2015)). The norms for news texts are also substantially different in these languages, making them more difficult to translate than texts in other domains, such as scientific paper abstracts (Nakazawa et al., 2019) and patent documents (Goto et al., 2013). The second reason is that we wished to conduct an in-depth analytic assessment of translation (see Section 5) by ourselves. We have a linguist who is highly competent in both linguistics and translation and has ample experiences in the analytic assessment of both MT outputs and human translations.

As material for this case study, we used the documents in the Asian Language Treebank (ALT) (Riza et al., 2016).[5] Table 1 gives statistics for the English source documents and Japanese target documents produced by professional human translators, where the numbers of tokens were counted after applying our in-house tokenizers.

[4] As a way of human evaluation, holistic assessment (or scoring) (Barrault et al., 2019; Nakazawa et al., 2019; Läubli et al., 2020; Barrault et al., 2020) is also beneficial, but does not suffice for our needs.

[5] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

Split	#Doc.	#Seg.	#Tok.	
			English	Japanese
Training	1,698	18,088	2,572k	3,743k
Development	98	1,000	139k	202k
Test	97	1,018	143k	208k

Table 1: Statistics for the ALT English–Japanese data (ALT-Standard-Split).

4 Data Collection

To clarify the limitations of the text-to-text approach for MT while acknowledging its status, we began with the outputs of a reasonably strong NMT system and collected examples of translation issues with their revisions through a modified version of the two-stage PE workflow originally proposed by Scarton et al. (2015). Our procedure is as follows.

Stage (1) Segment-level text-to-text NMT: Given source documents are translated by an MT system, which is preferably the one that can produce a translation of exploitable quality. We regard a segment-level text-to-text NMT as the subject.

Stage (2) Segment-level minimal PE: Each segment-level MT output is separately post-edited without referring to any information other than the segment itself, for example, other segments in the same document and other reference documents. To avoid introducing any preferences from human workers, this stage allows only minimal edits.

Stage (3) Document-level full PE: The results of stage (2) are further post-edited at document level to resolve the remaining issues caused by segment-level and/or text-to-text processing, where the human workers are allowed to refer to any necessary information. The resulting data exhibit the limitations of the segment-level text-to-text processing.[6]

Figure 1 compares our workflow (in the right-most path) with conventional human translation ("Non-MT workflow") and the prevalent one in TSPs ("MT+PE"), i.e., segment-level text-to-text MT followed by document-level manual full PE. Our workflow can be seen as an extension of "MT+PE" with an intermediate segment-level minimal PE stage.

The division of segment-level and document-level PE was originally proposed by Scarton et al. (2015) as a means of manually assessing the outputs of statistical MT (SMT) systems. Note that our subject is not the gap between segment-level and document-level text-to-text processing, i.e., MT systems, as in Scarton et al. (2015), but the limitation of such text-to-text processing. We therefore need to collect translation issues that can only be resolved by referring to information other than the given textual information. To exclude issues that can be resolved by referring only to the given textual information as much as possible, we decided to obtain translations that are attainable but closest to the outputs of text-to-text MT through minimal PE; we explicitly constrain the human workers by (i) prohibiting them from referring to any information other than the textual information and (ii) allowing only minimal edits,[7] while also avoiding subjective stylistic changes.[8] Even though document-level text-to-text MT

[6]Translation obtainable through this method is not necessarily of high quality because it is, in the end, *post-editese* (Toral, 2019). We plan to analyze the gap between PE-based translation and high-quality human translation, i.e., the art of translation, in our future work.

[7]This might be comparable with the goal of light PE (ISO/TC37, 2017): "obtain a merely comprehensible text without any attempt to produce a product comparable to a product obtained by human translation."

[8]Scarton et al. (2015) regarded style changes as the translator's choice. However, according to ISO/TC37 (2015), the appropriate style is not determined by the translators, but by the extra-document specifications for translation, for instance in the form of a translation brief that specifies the purpose/usage of the translated documents.

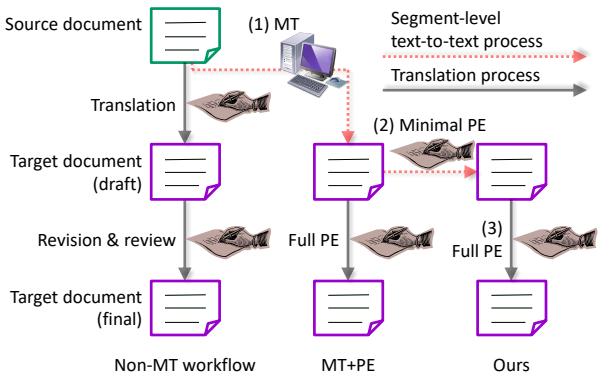

Figure 1: Comparison of translation workflows: the translation process refers to any information other than the given source document (cf. text-to-text process).

has been actively studied (Voita et al., 2018, 2019; Lopes et al., 2020), we decided to begin with segment-level MT and PE because we can ensure the minimality of the edits using segment-level automatic metrics (see Section 4.2).

By performing only minimal PE at segment level, we can leave all the translation issues that can only be resolved by referring to extra-document and/or non-linguistic information for a later stage. These issues are resolved in the succeeding document-level full PE stage, and we distinguish (a) those issues revealing the gap between segment-level and document-level processing and (b) those issues revealing the limitations of the text-to-text processing, through our manual analysis (see Section 5).

Our process for collecting translation issues uses some parameters that differ from those in Scarton et al. (2015), including the MT paradigm (SMT vs. NMT), translation task (English-to-German vs. English-to-Japanese), and worker experiences (students vs. professionals employed by a TSP with ISO certificates (ISO/TC37, 2015, 2017)).

4.1 Stage (1) Segment-level Text-to-Text NMT

To begin with a translation of exploitable quality, we trained a segment-level but reasonably strong[9] English-to-Japanese NMT system on a large-scale in-house English–Japanese parallel corpus (henceforth, *TexTra*)[10] in addition to the ALT training data, using a method for domain adaptation (Chu et al., 2017). First, we trained an English-to-Japanese NMT model on *TexTra* alone, explicitly excluding all the segment pairs in the ALT. For each source and target language, a sub-word vocabulary was also created from the corresponding side of this corpus: we determined 32k sub-words with byte-pair encoding (Sennrich et al., 2016b) after tokenization. Then, we fine-tuned the model parameters on a mixture of *TexTra* and the ALT training data. Following Chu et al. (2017), we used a balanced mixture of the two corpora by inflating the ALT training data K times and randomly sampling the same number of segment pairs from *TexTra*. Finally, we further fine-tuned the NMT model on the ALT training data only.

We used Marian NMT (Junczys-Dowmunt et al., 2018)[11] for all the NMT training and decoding processes, using the Transformer Base model and the hyper-parameters for training as

[9] We are aware that our system would not be state of the art because we do not use synthetic parallel data, a model ensemble, nor re-ranking. However, because these are all the methods for improving segment-level text-to-text MT, we assume that omitting them does not affect the main issues that we identify during the document-level full PE stage.

[10] The size is confidential. The generic model can be used via https://mt-auto-minhon-mlt.ucri.jgn-x.jp.

[11] https://github.com/marian-nmt/marian/, version 1.7.0

used in Vaswani et al. (2017). We terminated the training at each phase by early-stopping with a patience of 5, regarding the model perplexity on the ALT development data, computed after every T iterations, as the evaluation criterion. The value of T was set to 5,000 for the phase 1, and 10 for the phases 2 and 3. For the value of sample size K in phase 2, we selected 32 from the options 1, 2, 4, 8, 16, 32, and 64 according to the BLEU score (Papineni et al., 2002) on the ALT development data, computed by SacreBLEU (Post, 2018).[12] When decoding the ALT test data, the beam size was fixed 10, and the value for the length penalty was tuned on the ALT development data and set to 0.8.

4.2 Stage (2) Segment-level Minimal PE

To perform a segment-level PE, we isolated each segment from the others in the same document by shuffling the pairs of source segment and corresponding segment-level MT output across all the test documents.

We then asked[13] an experienced, ISO-certified TSP with well-designed workflows for translation (ISO/TC37, 2015) and PE (ISO/TC37, 2017) to revise the MT output of each segment independently without referring to any information other than the individual segment. The goal of this stage was to obtain a segment-level translation that fluently and accurately conveys the information in the corresponding source segment. To avoid excessive PE, we imposed a constraint, $hter(m, p) \leq hter(m, r)$, where m, p, and r stand for the MT output, its post-edited version, and reference translation,[14] respectively. $hter(a, b)$ is the Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), which computes how one segment a is dissimilar from another segment b at surface level, implemented in tercom.[15] We used MeCab[16] to tokenize the Japanese translation, unlike our implementation of NMT, in order to enable the consistent tokenization in both our environment and the workers' environment.

During this process, 95% of the segments (970/1,018) received some revisions. This suggests that our system still seldom generates acceptable segment-level translation in this English-to-Japanese news translation task. Because we allowed only minimal editing operations, the results represent the closest goal of segment-level text-to-text MT.

4.3 Stage (3) Document-level Full PE

After completing segment-level minimal PE for all segments, the documents were reverted by ordering the segments. We then asked[17] another set of workers through the same TSP to further revise the translation referring not only to the entire document but also to any extra-document and/or non-linguistic information, as in the ordinary document-level full PE workflow, i.e., "MT+PE" in Figure 1. Note that we hid the original MT outputs and provided the results of segment-level PE as the draft translation for revision. The workers were asked to make the target documents cohesive, consistent, and appropriate for news articles, also correcting content errors if any. Some examples are presented in Section 5.1.

As a result, 320 segments (31%) in 86 documents (89%) were revised. The total quantity of edits during this stage was much smaller than in the previous stage, but they were indeed necessary to obtain proper translations. This also confirms that a sequence of acceptable segment-level text-to-text translations does not necessarily qualify as translation. It further confirms that,

[12] https://github.com/mjpost/sacreBLEU/, short signature: BLEU+c.mixed+l.en-ja+#.1+s.exp+t.13a+v.1.4.1

[13] The price was based on the number of tokens in the source documents as in an ordinary translation contract. Thus, there was no incentive to increase the amount of PE.

[14] The TSP and workers did not see the ALT reference translation, and were asked to redo the task from the given MT output if we judged that their PE result did not satisfy the constraint.

[15] http://www.cs.umd.edu/~snover/tercom/, version 0.7.25.

[16] https://taku910.github.io/mecab/, version 0.996.

[17] For this task, we paid the same amount as we did for the segment-level PE.

Translation	BLEU (↑)	HTER (↓)
Output of NMT trained only on ALT	14.6	73.9
Output of NMT in phase 1	29.0	55.5
Output of NMT in phase 2	$35.8^{\dagger 1}$	47.6
Output of NMT in phase 3	$36.0^{\dagger 1}$	47.6
Segment-level minimal PE result	$36.8^{\dagger 3}$	47.0
Document-level full PE result	$36.8^{\dagger 3}$	47.0

Table 2: BLEU and HTER scores of different versions of translations with respect to the ALT reference translation (ALT). Note that these results are based on our in-house Japanese tokenizer (cf. MeCab used in the workflow for consistent tokenization). "†1" and "†3" respectively denote the score is significantly better than that for phases 1 and 3 ($p < 0.05$).

as in other well-studied translation tasks (Läubli et al., 2020; Freitag et al., 2021), *human parity* (Hassan et al., 2018) is not yet attainable in this English-to-Japanese news translation task.

4.4 Translation Quality Measured by Automatic Evaluation Metrics

Table 2 summarizes the BLEU and HTER scores of different versions of translations obtained in our workflow. To determine if differences in BLEU scores are significant, we performed statistical significance testing ($p < 0.05$).[18] The BLEU score of our adapted NMT system (phase 3) was significantly better than the non-adapted system (phase 1). We consider that it generated a translation of sufficient quality for this first stage in the process. Whereas the improvement brought by segment-level minimal PE was visible and the BLEU gain was statistically significant, the document-level full PE improved neither BLEU nor HTER scores.

5 Manual Analysis of Translation Issues

Our post-edited translation data contain two separate and different types of translation issues: the remaining issues from the segment-level text-to-text MT, and the issues that require information other than the individual segments to resolve. We manually analyzed the latter translation issues resolved during the document-level full PE in stage (3).

First, using tercom, we automatically identified the corresponding text spans in the two versions of the translations obtained in stages (2) and (3). Then, we manually extracted pairs of text spans: one for an issue in the segment-level PE result, and the other for its revision in the document-level PE result. As a result, we obtained 529 such *revision examples*. Finally, we annotated each revision example with the following three types of labels.

Need for document-level textual information: whether the textual information outside the segment but within the document was necessary to solve the issue.

Need for extra information: whether any extra-document and/or non-linguistic information was necessary to solve the issue. If it was needed, we also noted the information types (more than one if applicable).

Issue type: one of the 16 types in a translation issues typology designed for assessing and learning English-to-Japanese translation (Fujita et al., 2017). We chose this typology because its usefulness for this translation direction had been verified, whereas a widely used MQM had not.

[18] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/analysis/bootstrap-hypothesis-difference-significance.pl

		Extra info.	
		No need	Necessary
Document-level textual info.	No need	(a) 196	(c) 168
	Necessary	(b) 116	(d) 49

Table 3: Revision examples classified according to the types of necessary information.

Issue type		#Examples		
		(a)	(b)	(c),(d)
Lv 1: **Incompleteness**	X4a: Content-untranslated	0	0	16
	X6: Content-indecision	0	1	1
Lv 2: **Semantic errors**	X7: Lexis-incorrect-term	5	6	67
	X1: Content-omission	11	4	2
	X2: Content-addition	3	0	0
	X3: Content-distortion	42	31	25
Lv 3: **Linguistic issues in target document**	X8: Lexis-inappropriate-collocation	5	0	0
	X10: Grammar-preposition/particle	2	0	0
	X11: Grammar-inflection	0	0	0
	X12: Grammar-spelling	0	0	0
	X13: Grammar-punctuation	7	0	0
	X9: Grammar-others	1	0	0
Lv 4: **Felicity issues in target document**	X16: Text-incohesive	31	63	14
	X4b: Content-too-literal	54	0	7
	X15: Text-clumsy	35	3	4
Lv 5: **Register issues in target document**	X14: Text-TD-inappropriate-register	0	8	81
Total		196	116	217

Table 4: Distribution of the revision examples. Refer to Fujita et al. (2017) for the definition of each issue type and the classification procedure, and Table 3 for the classification of (a) to (d).

Tables 3 and 4 show our classification results: whereas Table 3 shows a contingency table based on the first two labels, Table 4 shows the type-wise numbers of revision examples, merging (c) and (d) in Table 3 for the sake of simplicity.

5.1 Issues Beyond Text-to-text MT

Among the four classes shown in Table 3, our main subjects are 217 examples in (c) and (d) that can only be resolved by referring to some extra-document and/or non-linguistic information. Such information is categorized into the following four types.

A) Fine-grained style specifications (121 examples): Texts in Japanese newspapers are written following various specifications, including those for vocabulary, set of characters, usages of symbols including parentheses, degree of formality, and other notational rules. Our source texts themselves might have revealed that they are from the news domain. However, the workers for the segment-level minimal PE task did not perform revisions to fulfill such specifications, leading to translation that is inappropriate for the register (81 X14 issues). Because of a lack of a specification for transliteration at the segment-level PE stage, the workers left some named entities untranslated (16 X4a issues), considering that Latin characters are sometimes used in Japanese documents and that the contents in the source segments are comprehensible.

Source:	Clemens (3-0(#1), 1.90 ERA in seven World Series starts) will make his 33rd career postseason(#6,#7) start(#8) Saturday, at least for a day matching(#5) Pettitte (3-4(#3), 3.90 in 10 World Series starts) for the most ever(#4).
Seg.PE:	クレメンス（ワールドシリーズ７回出場で３対０(#1/3 points vs 0 points)、 防御率１．９０）は、少なくとも１日 [ ](#2/ϵ)ペティット（ワールドシリーズ１０回出場で ３対４(#3/3 points vs 4 points)、 ３．９０）と [ ](#4/ϵ)組んで(#5/paired)、 土曜日に [ ](#6/ϵ)３３回目のポストシーズン(#7/33rd postseason)のスタートを切る(#8/start)。
Doc.PE:	クレメンス（ワールドシリーズ７回出場で３勝０敗(#1/3 wins and 0 losses/(c)/X3)、 防御率１．９０）は、少なくとも１日 は(#2/topic marker/(a)/X15) ペティット（ワールドシリーズ１０回出場で３勝４敗(#3/3 wins and 4 losses/(c)/X3)、 ３．９０）と史上最多で(#4/most ever/(a)/X1)並び(#5/ranked same/(c)/X3)、 土 曜 日 に生涯で(#6/in ones life/(a)/X1)ポストシーズン３３回目(#7/33rd time in postseason/(c)/X3)の先発登板を行う(#8/to be the first pitcher of the game/(c)/X3)。

Figure 2: An example segment (Doc.ID: 24312, Seg.ID: 15534), where eight issues (numbered in the first element of subscript) were resolved during the document-level full PE. The second elements of the subscript in the translation give phrase-level gloss, and the remaining elements of the subscript for the document-level full PE represent the type of necessary information (see Table 3) and the issue type (see Table 4).

B) Terminology (80 examples): When translating named entities, we must look up the terminologies for authorized translations/transliterations. Consider, for instance, the person name "John Paul." The most likely transliteration for it is "ジョン・ポール" (/dʒ'ɔn p'ɔ:l/). However, it must be transliterated into "ヨハネ・パウロ" (/johʌnɛ pʌʊlə/) when it refers to the Pope. Most improper and/or inconsistent term translations (64 X7 issues) and the above untranslated entities (16 X4a issues) were caused due to a lack of a terminology.

C) Domain-specific knowledge (31 examples): Our documents cover diverse topics such as politics, religion, and sports. Some semantic issues required knowledge specific to each of these domains to understand the contents in the source texts and produce appropriate expressions. See, for instance, the example in Figure 2. One must realize that this text is talking about baseball, and have knowledge about that domain, in order to perform the revisions marked (c). Some incohesive issues (five X16 issues) also require such knowledge to resolve.

D) Reference documents (eight examples): When translating ambiguous expressions, we need some clues to disambiguate them. If the document does not contain such information, we must find some reliable information outside the document. Because our text-to-text MT system and our segment-level minimal PE can only access the textual information, some semantic issues (seven X3 issues) and an incomplete translation with multiple options (X6 issue) were left. The X6 issue gives both "兄 (elder brother)" and "弟 (younger brother)" as multiple translation options for "brother." This ambiguity was resolved only when the worker found credible biographical information on the Web. Although we found only eight examples that were resolved in the document-level full PE stage referring to other information sources, we confirmed that our text-to-text MT sometimes correctly disambiguates such expressions by chance.

5.2 Remaining Issues of Text-to-Text MT

The remaining 196 and 116 examples were respectively classified as (a) and (b), i.e., those that had been resolved by referring only to the given textual information. These resolutions could be attainable by algorithmic advancements in the text-to-text approach for MT. Although they are outside the focus of this paper, we make some observations relevant to our study.

Segment-level issues, i.e., (a), lie at the levels 2 to 4 in the issue typology (Table 4). Whereas the ones at levels 2 and 3 should have been resolved through segment-level minimum PE, the ones at level 4 are not considered mandatory as long as the translations are considered comprehensible. We believe that we have successfully excluded much larger number of similar segment-level text-to-text issues by introducing the segment-level minimal PE stage (Section 4.2) and the above remaining issues are not harmful to our study. We could have reduced the examples in this class by removing our constraints for minimal edits. However, this introduces some risks, such as losing examples in our concern, i.e., (c) and (d), and being mislead by some artificial examples, such as combinations of preferential edits in both segment-level PE and document-level PE.

Class (b) examples exhibit revisions made by referring to the textual information in the document, but no more than that. They appeared at all issue levels in the typology except level 3, grammaticality, and the majority were either X16 (incohesive) or X3 (content distortion). To translate the mentions of each entity coherently and cohesively (Voita et al., 2019), we need to identify the correct referent of each mention. In the literature, a matrix called the *entity grid* (Barzilay and Lapata, 2008) is used to represent the appearance of entities and segments in the given source document. Actively studied document-level text-to-text MT might be able to capture such information, for instance, by enhancing the self-attention mechanisms (Vaswani et al., 2017; Maruf et al., 2019; Beltagy et al., 2020). However, as we confirmed in our analysis (Section 5.1), referents are not necessarily given in the source document, and we hence must seek reliable extra-document information.

6 Discussion and Future Directions

Techniques for MT have been advanced thanks to the simplified problem setting, i.e., text-to-text processing, and the advent of automatic evaluation metrics, such as BLEU (Papineni et al., 2002), which are based on comparison with reference translations. However, considering the large gap between what text-to-text MT can ultimately attain and the needs that translation must satisfy, a fully automatic MT approach (Hutchins and Somers, 1992) still looks infeasible. Rather, approaches in machine-aided human translation and human-aided MT, i.e., human–machine interactions, are more promising. Indeed, "MT+PE" in Figure 1, which has been prevalent in the translation production workflow at TSPs for a decade, lies in that direction. In this way, to reduce the cognitive load of PE, we must continue to enhance both wheels, i.e., improving MT systems and determining the best practices in using them.

As confirmed in Section 4.2, segment-level text-to-text MT still has much room for improvement. Yet, as shown in recent studies, textual information within the entire source document is useful. To generate cohesive texts, we should incorporate the latest outcomes in discourse processing and natural language generation, such as discourse parsing (Jia et al., 2018) and generating referential expressions (Paraboni et al., 2007). To assess MT outputs for further improvement while reducing the human labor in PE, we also need to invent document-level automatic evaluation methods, preferably analytic ones rather than holistic ones. Ultimately and ideally, we should also consider going beyond text-to-text processing, seeking better ways for incorporating information indispensable for translation, such as those we described in Section 5.1, rather than indirectly representing them with text data. For instance, to enforce the use of particular expressions specified by pre-compiled terminologies and style specifications, we need to improve the decoding mechanism, such as constrained decoding (Hasler et al., 2018; Post and Vilar, 2018; Zhang et al., 2018). Style specifications and domain-specific knowledge might be learned from text data in a given fine-grained domain, such as the one in Figure 2. We can see related work in adaptive data selection (Chen et al., 2016) and extreme adaptation (Michel and Neubig, 2018a).

In addition to the enhancement of MT systems, we should also establish reliable and effective ways for identifying critical issues in MT outputs as well as determining translation scenarios where MT is promising or hopeless. For instance, word frequency and sentence length affect the segment-level MT quality (Koehn and Knowles, 2017). Such findings motivate the *pre-editing* of segments prior to decoding (Pym, 1990; Miyata and Fujita, 2021).

From a general perspective, we should consider educating people (all people) so that they acquire two types of literacy: *translation literacy* for understanding the norms, skopos, and other specifications in their translation task (Klitgård, 2018), and *MT literacy* for understanding the characteristics of the intended MT service, which helps minimize potential risks (Bowker and Ciro, 2019). We believe that our method for clearly delineating between translation and the translation that text-to-text MT can ultimately attain as well as our case-study findings can be useful resources for such education.

7 Conclusion

To analytically assess issues that cannot be resolved by text-to-text processing, such as text-to-text MT, this paper presented our specific constraints incorporated into the two-stage PE pipeline originally proposed by Scarton et al. (2015). In a case study on the English-to-Japanese news translation task, we found that translation issues beyond text-to-text processing are caused by a lack of extra-document and/or non-linguistic information, such as fine-grained style specifications, terminology, domain-specific knowledge, and reference documents. The resulted parallel data and annotated revision examples are publicly available.[19]

Our method is laborious and requires very high competence in both linguistics and translation. Nevertheless, it is applicable to other translation tasks where we can build an MT system that can produce translation of exploitable quality. We thus hope other researchers use our method to assess the limitations of text-to-text processing and the remaining issues in a wide range of translation tasks. We plan to introduce another document-level minimal PE stage in order to assess the attainable translation by document-level MT.

While clarifying the limitations, we also suggested how we can enable MT systems to explicitly refer to extra-document and/or non-linguistic information. We plan to evaluate the impact of enforcing decoding with external knowledge, such as terminologies and style specifications.

An important issue in present-day society was also illuminated: the need to cultivate translation literacy and MT literacy in people to avoid the risk caused by the innocent use of MT services. To tackle this, we are currently compiling educational materials to help people understand translation, MT, and their differences. We will also analyze various levels of competences required for human translators, following the Competence Framework developed by the European Master's in Translation (Toudic and Krause, 2017).

Acknowledgments

I am deeply grateful to Masao Utiyama for giving me permission to use *TexTra* as well as Kyo Kageura, Masaru Yamada, Rei Miyata, Takuya Miyauchi, and Mayuka Yamamoto for their insightful comments on translation revisions. I would also like to thank Raj Dabre, Hideki Tanaka, and the anonymous reviewers, including those for past submissions, for their valuable comments on earlier versions of this paper. This work was partly supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) 19H05660.

[19] https://github.com/akfujita/staged-PE

References

Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 1557–1567.

Bapna, A. and Firat, O. (2019). Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), pages 1538–1548.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., kiu Lo, C., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In *Proceedings of the 3rd Conference on Machine Translation* (*WMT*)*: Shared Task Papers*, pages 304–323.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document Transformer. *CoRR*, abs/2004.05150.

Bowker, L. and Ciro, J. B. (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Group Publishing Ltd.

Castilho, S., Popović, M., and Way, A. (2020). On context span needed for machine translation evaluation. In *Proceedings of the 12th International Conference on Language Resources and Evaluation* (*LREC*), pages 3735–3742.

Chatterjee, R., Negri, M., Turchi, M., Federico, M., Specia, L., and Blain, F. (2017). Guiding neural machine translation decoding with external knowledge. In *Proceedings of the 2nd Conference on Machine Translation* (*WMT*), pages 157–168.

Chen, B., Kuhn, R., Foster, G., Cherry, C., and Huang, F. (2016). Bilingual methods for adaptive training data selection for machine translation. In *Proceedings of the 12th Biennial Conference of the Association for Machine Translation in the Americas* (*AMTA*), pages 93–106.

Chesterman, A. (1997). *Memes of Translation: The Spread of Ideas in Translation Theory*. John Benjamins.

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*ACL*)*, Short Papers*, pages 385–391.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *CoRR*, abs/2104.14478.

Fujita, A., Tanabe, K., Toyoshima, C., Yamamoto, M., Kageura, K., and Hartley, A. (2017). Consistent classification of translation revisions: A case study of English–Japanese student translations. In *Proceedings of the 11th Linguistic Annotation Workshop* (*LAW*), pages 57–66.

Goto, I., Chow, K. P., Lu, B., Sumita, E., and Tsou, B. K. (2013). Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, pages 260–286.

Hardmeier, C. (2014). *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University.

Hashimoto, K., Buschiazzo, R., Bradbury, J., Marshall, T., Socher, R., and Xiong, C. (2019). A high-quality multilingual dataset for structured documentation translation. In *Proceedings of the 4th Conference on Machine Translation* (*WMT*) *(Volume 1: Research Papers)*, pages 116–127.

Hasler, E., de Gispert, A., Iglesias, G., and Byrne, B. (2018). Neural machine translation decoding with terminology constraints. In *Proceedings of Human Language Technologies: The 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 506–512.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

Irvine, A., Morgan, J., Carpuat, M., III, H. D., and Munteanu, D. (2013). Measuring machine translation errors in new domains. *Transaction of the Association for Computational Linguistics* (*TACL*), 1:429–440.

ISO/TC37 (2015). ISO 17100:2015 translation services: Requirements for translation services.

ISO/TC37 (2017). ISO 18587:2017 translation services: Post-editing of machine translation output: Requirements.

Jia, Y., Ye, Y., Feng, Y., Lai, Y., Yan, R., and Zhao, D. (2018). Modeling discourse cohesion for discourse parsing via memory network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*ACL*)*, Short Papers*, pages 438–443.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*ACL*)*, System Demonstrations*, pages 116–121.

Kenny, D. (2001). *Lexis and Creativity in Translation: A Corpus Based Approach*. Routledge.

Klitgård, I. (2018). Calling for translation literacy: The use of covert translation in student academic writing in higher education. *Translation and Translanguaging in Multilingual Contexts*, 4(2):306–323.

Kobus, C., Crego, J., and Senellart, J. (2017). Domain control for neural machine translation. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 372–378.

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

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., and Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Lommel, A., Görög, A., Melby, A., Uszkoreit, H., Burchardt, A., and Popović, M. (2015). QT21 deliverable 3.1: Harmonised metric.

Lopes, A. V., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. T. (2020). Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association of Machine Translation* (*EAMT*), pages 225–234.

Maruf, S., Martins, A. F. T., and Haffari, G. (2019). Selective attention for context-aware neural machine translation. In *Proceedings of Human Language Technologies: The 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 3092–3102.

Melby, A. K. (2012). Structured specifications and translation parameters (version 6.0). http://www.ttt.org/specs/.

Michel, P. and Neubig, G. (2018a). Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*ACL*)*, Short Papers*, pages 312–318.

Michel, P. and Neubig, G. (2018b). MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 543–553.

Miyata, R. and Fujita, A. (2021). Understanding pre-editing for black-box neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL*), pages 1539–1550.

Moussallem, D., Ngomo, A.-C. N., Buitelaar, P., and Arcan, M. (2019). Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of the 10th International Conference on Knowledge Capture* (*K-CAP*), page 139–146.

Nakazawa, T., Doi, N., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., Bojar, O., and Kurohashi, S. (2019). Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation* (*WAT*), pages 1–35.

Nida, E. A. (1964). *Toward a Science of Translating*. Brill.

Niu, X., Martindale, M., and Carpuat, M. (2017). A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 2814–2819.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 311–318.

Paraboni, I., van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254.

Popović, M. and Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the 3rd Conference on Machine Translation* (*WMT*)*: Research Papers*, pages 186–191.

Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of Human Language Technologies: The 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 1314–1324.

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

Riza, H., Purwoadi, M., Gunarso, Uliniansyah, T., Ti, A. A., Aljunied, S. M., Mai, L. C., Thang, V. T., Thai, N. P., Chea, V., Sun, R., Sam, S., Seng, S., Soe, K. M., Nwet, K. T., Utiyama, M., and Ding, C. (2016). Introduction of the Asian Language Treebank. In *Proceedings of the 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*, pages 1–6.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015). Searching for context: a study on document-level labels for translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association of Machine Translation* (*EAMT*), pages 121–128.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Controlling politeness in neural machine translation via side constraints. In *Proceedings of Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 35–40.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 1715–1725.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas* (*AMTA*), pages 223–231.

Toral, A. (2019). Post-editese: an exacerbated translationese. In *Proceedings of the 17th Machine Translation Summit* (*MT Summit XVII*), pages 273–281.

Toral, A. (2020). Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association of Machine Translation* (*EAMT*), pages 185–194.

Toudic, D. and Krause, A. (2017). European Master's in translation: EMT competence framework.

Toury, G. (1978). The nature and role of norms in literary translation. In Holmes, J., Lambert, J., and van den Broeck, R., editors, *Literature and Translation*, pages 83–100. Levine.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems* (*NIPS*), pages 5998–6008.

Vermeer, H. J. (1992). Is translation a linguistic or a cultural process? *Ilha do Desterro*, 28:37–49.

Vermeer, H. J. (2004). Skopos and commission in translational action. In Venuti, L., editor, *The Translation Studies Reader*, pages 227–238. Routledge.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 1198–1212.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 1264–1274.

Ye, Y. and Toral, A. (2020). Fine-grained human evaluation of transformer and recurrent approaches to neural machine translation for English-to-Chinese. In *Proceedings of the 22nd Annual Conference of the European Association of Machine Translation* (*EAMT*), pages 125–134.

Zhang, J., Utiyama, M., Sumita, E., Neubig, G., and Nakamura, S. (2018). Guiding neural machine translation with retrieved translation pieces. In *Proceedings of Human Language Technologies: The 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 1325–1335.