Fine-Grained Error Analysis on English-to-Japanese Machine Translation in the Medical Domain

Takeshi Hayakawa
Graduate School of Information Science and Technology,
Osaka University, Osaka, Japan
ASCA Corporation, Osaka, Japan
hayakawa.takeshi@ist.osaka-u.ac.jp

Yuki Arase
Graduate School of Information Science and Technology,
Osaka Un iversity, Osaka, Japan
arase@ist.osaka-u.ac.jp

Abstract

We performed a detailed error analysis in domain-specific neural machine translation (NMT) for the English and Japanese language pair with fine-grained manual annotation. Despite its importance for advancing NMT technologies, research on the performance of domain-specific NMT and non-European languages has been limited. In this study, we designed an error typology based on the error types that were typically generated by NMT systems and might cause significant impact in technical translations: "Addition," "Omission," "Mistranslation," "Grammar," and "Terminology." The error annotation was targeted to the medical domain and was performed by experienced professional translators specialized in medicine under careful quality control. The annotation detected $4,912$ errors on $2,480$ sentences, and the frequency and distribution of errors were analyzed. We found that the major errors in NMT were "Mistranslation" and "Terminology" rather than "Addition" and "Omission," which have been reported as typical problems of NMT. Interestingly, more errors occurred in documents for professionals compared with those for the general public. The results of our annotation work will be published as a parallel corpus with error labels, which are expected to contribute to developing better NMT models, automatic evaluation metrics, and quality estimation models.

1 Introduction

We performed a manual annotation of translation errors using fine-grained error typology in domain-specific neural machine translation (NMT) of Japanese and English language pairs. Although several approaches have been proposed to evaluate the performance of NMT, it has been commonly presented as scores of automatic evaluation, and detailed analysis of problems in NMT is limited. Previous studies (Specia et al., 2017; Kepler et al., 2019) annotated errors in MT outputs; however, they targeted only on a general domain and European languages. Detailed error detection is essential, especially in the domain-specific settings, where tiny mistakes, such as incorrect translation of a technical term, leads to significant misunderstanding.

To tackle this problem, we performed an annotation-based analysis of errors that occurred in NMT for a specific technical domain. Professional translators annotated types and positions of errors that occurred in translation from English to Japanese. The error typology was designed based on an existing framework, Multidimensional Quality Metrics (MQM) (Lommel et al., 2014), which was customized to our study. We selected medicine as the domain field because medical translation is in growing demand in the society to enrich healthcare information, which requires highly specific domain expertise. Recent issues regarding public health, such as the pandemic

of coronavirus disease 2019, highlight demands on sharing correct and understandable information throughout the world including Asian countries. We prepared five medical contents with English-to-Japanese translation data using state-of-the-art NMT systems. As a result, 4,912 errors in five types were annotated on 2,480 sentences. We also analyzed the annotation results in detail to reveal distributions and characteristics of errors produced by current NMT systems.

The results of annotation will be published as a parallel corpus with error labels. This is the first corpus of error annotation (1) on domain-specific and (2) on English-to-Japanese NMT outputs. Such corpora annotating errors in machine translation (MT) are valuable resources to understand problems in NMT models, develop automatic evaluation metrics, and estimate the quality of machine translation (Blatz et al., 2004).

2 Related Work

Our annotation corpus is based on the error typology that conforms to structured categories of quality metrics for translation quality. Previous studies employed a few different typologies, such as MQM and SCATE (Smart Computer-aided Translation Environment) (Tezcan et al., 2017). Among them, MQM is one of the most common frameworks for quality assessment of human translation. The framework of the typology in our study also refers to the MQM.

QT21 Consortium has published post edited and error annotated data for machine translations in four languages: Czech, English, German, and Latvian (Specia et al., 2017) based on MQM. This data just included languages in Europe, and prior studies that used the MQM have evaluated translation of European languages (Klubička et al., 2018; Van Brussel et al., 2018). Our corpus in English to Japanese will add a useful resource of annotation. The shared task of quality estimation in the Conference on Machine Translation (WMT) has also employed the MQM for document-level quality estimation since 2018. Approaches of quality estimation tasks with MQM include word-level annotation (Specia et al., 2018) and the estimation of MQM score with prediction models (Kepler et al., 2019). Nonetheless, there has been a limited resource for domain-specific translation (Rigouts Terryn et al., 2019), which is indispensable to develop an evaluation strategy for appropriateness of word choice in the technical context.

3 Error Typology & Development of Annotation Guidelines

In this study, we developed customized error-typology criteria for the evaluation of domain-specific NMT. Our typology was based on MQM. The major error categories in MQM are "Accuracy," "Fluency," "Design," "Locale convention," "Style," "Terminology," and "Verity," of which subcategories are defined for a specific type of incorrectness.

We selected and customized several error subtypes in the original MQM for annotation that were applicable to translations by NMT systems. In this paper, we focused on subtypes that annotation results confirmed as the major problems of the current NMT systems, namely, "Addition," "Omission," and "Mistranslation" from "Accuracy;" "Terminology;" and "Grammar" from "Fluency;" as summarized in Table 1.

We customized these error subtypes to handle domain specificity and the Japanese language due to different systems of grammar and sociolinguistic register from Western languages. The following sections describe these error types and guidelines given to annotators to identify each error.[1]

3.1 Addition and Omission

Over- and under-generations are typical errors in NMT because of the lack of a mechanism to explicitly track source-sentence coverage (Tu et al., 2016). These were categorized as "Addition" and "Omission," respectively.

"Addition" and "Omission" errors occur only in target and source sentences, respectively. Our guidelines instructed annotators to assign a label of "Addition" on the word(s) of target sentence that does not semantically correspond to any word in the source sentence. On the contrary, the guidelines required to attach a label of "Omission" to the word(s) of

[1] The guidelines are attached to our corpus to be released.

Error type	Description of error	Annotation span	Annotation side
Addition	The target text includes text not present in the source.*	Word/Phrase	Target
Omission	Content is missing from the translation that is present in the source.*	Word/Phrase	Source
Mistranslation	The target content does not accurately represent the source content.*	Word/Phrase	Source
Terminology	The target text is not suitable in terms of the domain of document.	Word/Phrase	Source
Grammar	Syntax or function words are presented incorrectly.	Word/Phrase	Target

Table 1: Error typology (Descriptions with asterisks are cited from MQM Issue Types.)

the source sentence of which translation did not appear in the target sentence. In cases that grammatical words specific to the target language were not translated, this kind of errors was not considered as "Omission" but as "Grammar."

Relevant error subtypes to "Addition" and "Omission" defined in MQM are "Over-translation" and "Under-translation." These apply to a translation output that is more or less specific than the source sentence, respectively. Different from human translation, our annotation results revealed that Over- and Under-translations were far infrequent in current NMT systems.

3.2 Mistranslation

This type of error refers to the semantic difference between words or phrases in source and target sentences. The wrong choice of meaning in polysemous words was included in the "Mistranslation," as well as incorrect translation.

The guidelines instructed annotators to assign a label of "Mistranslation" on the word(s) of a source sentence that was incorrectly translated. We distinguished mistranslation and terminological errors to identify domain-specific errors. Hence, inappropriate use of words with the same or similar meaning in translation was categorized to "Mistranslation," as discussed below.

3.3 Terminology

We incorporated the appropriateness of word choice to our typology as the category of "Terminology," to ensure applicability to measure the domain specificity of translation outputs. We defined terminology errors as a translated word that was unsuitable to the description in the medical field, even though the meaning of the word was acceptable in the translation of the general domain.

The "Mistranslation" and "Terminology" errors were distinguished whether a translation output correctly reflected the meaning of the source sentence.

Our guidelines instructed annotators that the errors in the choice of technical terms with similar meaning should be labeled as "Terminology," instead of "Mistranslation." On the contrary, if a translated word(s) was semantically incorrect, the word was assigned the "Mistranslation" label, irrespective of the presence of "Terminology" error. The labels of "Terminology" were placed on the source sentence.

For example, the word "primary" means "most important" or "coming earliest" in general, but when used as "primary tumor" in the context of medicine, it means "the originally developed cancer cells in the body." Hence, translating "primary tumor" as "most important tumor" is regarded as "Terminology" error, while translating into "new tumor" is regarded as a "Mistranslation" error.

3.4 Grammar

Grammatical errors in English-to-Japanese translation affect the quality of translation more significantly. This is because grammatical errors in English-to-Japanese translation are characterized by incorrect understanding of syntax, which often changes the meaning of source sentence. For example, incorrect translation output of Japanese particles may be presented as the conversion between subjective and objective cases.

The guidelines instructed annotators to assign a label of "Grammar" on the target sentence for the errors of incorrect syntax representation, grammatically inappropriate output, and wrong order of words.

3.5 Sides of Annotation

The right-most column of Table 1 shows whether annotations were conducted on source sentences or translation outputs for each error type. Since MQM has not determined which side of the sentence the error should be labeled, in this study, we defined the annotation side specific to each error type. "Addition" and "Omission" were marked on target and source sides, respectively, because their occurrences are one-sided. As for "Mistranslation" and "Terminology," we attached the labels on only source sentences for simplicity of the annotation process. The alignment of these source words and phrases to the target-side is subject to our future work. The "Grammar" error was marked in the target-side because annotators can identify ungrammatical parts in a sentence, but it was hard to determine what caused these grammatical errors.

4 Annotation Setup

In this section, we describe the annotation procedure and resources used to perform the annotation.

4.1 Annotation Procedure

First of all, annotators were instructed to read through the annotation guidelines before starting the annotation and to be familiar with the standards. The annotators were provided triples of a source sentence, reference translation, and MT output, and worked for annotation through October to December 2018. The annotators identified spans of word/phrase/sentence presenting errors and assigned the corresponding error types as labels on the sentence level. Annotation could be overlapped on the same spans for different types of errors.

4.2 NMT Systems

Distribution of the occurrence of errors might depend on a certain translation system; therefore, we used multiple systems to reduce the effect of such dependency. We used state-of-the-art NMT systems for English-to-Japanese translation available in October 2018 at the time of annotation, as described below.

- Google' s neural machine translation system (GNMT) (Wu et al., 2016)
- NICT's neural machine translation system (Wang et al., 2018) (NICT NMT)

The preliminary investigation confirmed that there was no substantial difference between both systems. The corpus-level BLEU scores of GNMT and NICT NMT were 36.20 and 35.70, respectively. The mean normalized Levenshtein distance[2] of each sentence between references and translation outputs of GNMT and NICT NMT were 0.64 ($\pm$0.23) and 0.64 ($\pm$0.22), respectively. Paired bootstrap resampling test (Koehn, 2004) showed no significant difference in the two NMT systems for corpus BLEU ($p = 0.17$) as well as Student's t-test for normalized Levenshtein distance ($p = 0.63$); hence, we did not distinguish their outputs in the later processes.

4.3 Corpora for Annotation

Our annotation corpus consisted of 2,480 sentences from the medical/pharmaceutical domain in English. We collected the sentences from five sources of documents with different types: MSD Manual Consumer Version (Merck and Co., Inc., 2015a), MSD Manual Professional Version (Merck and Co., Inc., 2015c), New England Journal of Medicine (Massachusetts Medical Society, 2019), Journal of Clinical Oncology (American Society of Clinical Oncology, 2019), and ICH guidelines (Singh, 2015). Two versions of MSD manual are for the same topics of medical information but differentiated by expertise levels of contents: Professional Version includes highly technical terms for health professionals, and Consumer Version is written for the general population without domain knowledge.[3] New England Journal of Medicine and Journal of Clinical Oncology are standard academic journals of medicine. ICH guidelines consist of international regulations for pharmaceutical manufacturing processes. The source sentences were randomly extracted from each document.

We obtained the Japanese translation of the corpora from the two NMT systems. The set of target sentence was produced by randomly

[2] Levenshtein distance divided by the length of reference and target sentences.

[3] Therefore, the Consumer and Professional versions consist of comparable sentences with different expertise levels but are not exactly parallel.

Source	Expertise Level	Number of sentences	Mean number of words per sentence	BLEU	normalized Levenshtein distance
MSD Manual Consumer Version	General	580	17.88 (±7.89)	31.58	0.66 (±0.23)
MSD Manual Professional Version	Professional	560	19.50 (±9.48)	38.93	0.59 (±0.24)
New England Journal of Medicine	Professional	420	29.96 (±17.12)	37.65	0.62 (±0.21)
Journal of Clinical Oncology	Professional	420	22.99 (±12.09)	36.29	0.69 (±0.24)
ICH guidelines	Professional	500	18.08 (±5.77)	33.67	0.66 (±0.21)
Total		2,480	21.20 (±11.61)	35.95	0.64 (±0.23)

Table 2: Statistics of language resource for annotation

selecting each translated sentence from the two NMT outputs (50% for each), to prepare bilingual pairs of the 2,480 sentences. Table 2 shows the statistics of our annotation corpus.

These source sentences have corresponding Japanese versions, which were prepared by human translation with the professional review (Merck and Co., Inc., 2015d; Merck and Co., Inc., 2015b; Nankodo Co.,Ltd., 2019; American Society of Clinical Oncology, 2018; Pharmaceuticals and Medical Devices Agency, 2018). These Japanese versions were used as the reference translations.[4]

4.4 Annotators

To ensure the quality of annotation, we recruited three professional translators in the medical/pharmaceutical field. All the annotators were native Japanese translators with an academic background in biology or pharmacology. Year of translation experience ranged from three to eight years. The annotators identified errors and their types in an NMT output referring to corresponding source and reference translations.

5 Quality Control of Annotation

This kind of error annotation is inevitably subjective, because the ability to detect errors in translation depends on the level of expertise. In addition, determination of the type and span of errors should be contingent on the preference of each annotator, which may cause the variation of the annotation work.[5]

In this study, to collect reliable annotations alleviating such subjectivity, we conducted a pilot study and reconciliation of annotated labels.

5.1 Pilot Study

We performed a pilot study with the annotators using an independent data, consisted of 100 pairs of sentences.

Annotations on the pilot study were thoroughly reviewed by the authors and feedbacked to the annotators when there were misunderstandings of the guidelines. Also, questions raised by any annotator and the answers were shared to ensure that annotators have the same understanding of the task.

5.2 Reconciliation of Annotation

Once the annotators completed the annotation, they reviewed all the annotation results from the other annotators. They judged whether to accept or reject each annotation label. When two or more annotators voted to accept an annotation label, the corresponding annotation is retained, otherwise discarded.

The first annotation process identified 7,424 errors. The three annotators assigned 3,115 labels on average, with a standard deviation of 37.82. After the reconciliation process, the total number of errors with types was reduced to 4,912. Among these, 4,572 annotations were agreed by all the three annotators, and the rest 340 were agreed by two, which shows that our final annotation results are highly reliable. Note that 2,352 errors with the same labels and spans were consolidated as one error. Errors with overlapping span but with different labels were kept as independent annotations. Annotations on partially overlapping span with same error type were combined to one annotation that had larger span (e.g. Two annotations on “a condition” and “condition”

[4] Some of the Japanese articles in the MSD manual are comparable but not parallel translations because of difference in edition and local regulation. Therefore, we manually selected sentences ensuring the equivalence of the translation pairs.

[5] Due to this variation, a common metric to measure the agreement of annotations, i.e., Fleiss' Kappa, is not applicable.

were combined to that on "a condition.").

We confirmed that "Terminology," "Addition," and "Omission" errors were highly agreed (96.8%, 71.4%, and 64.1% of errors were accepted by at least two annotators). On the other hand, "Mistranslation" and "Grammar" errors had an opposite tendency (46.0% and 47.4% were accepted by at least two annotators). The disagreement of annotation separating "Mistranslation" and "Terminology" was effectively combined through the reconciliation work. The judgment of "Mistranslation" and "Terminology" errors tended to be more subjective, which caused disagreement. These results imply that the many cases of disagreement were reconciled as "Terminology" error, rejecting the annotation of "Mistranslation." In addition, annotators commented that "Addition" and "Omission" errors were harder to detect and large part of disagreement in these errors were due to oversight. Therefore, the reconciliation resulted in the high acceptance ratios.

5.3 Annotation Examples

Table 3 shows examples of annotation results after reconciliation, in which underlined phrases in the text indicate errors. The first case is an example of "Addition," in which the same words of "長期的な (long-term)" appear twice in the target sentence. The second appearance was annotated as "Addition." In the second case, the translation corresponding to the words "both of" in the source sentence is not included in the target sentence. This type of error was annotated as "Omission." The third and fourth cases represented "Terminology" errors. In the third case, the word "at 90 days" was used to mean a time point; however, the MT output referred to duration, and thus annotated as "Mistranslation." In the fourth case, "may" was used to express a possibility, which was not reflected in the target output. The fifth case is an example of "Grammar." In this case, the coordination in the source sentence means "low vitamin D intake or low calcium intake;" however, the translation in the target text means "low vitamin D, and calcium intake." This type of syntax error was annotated as "Grammar." The sixth and seventh cases represented "Terminology" errors. In the sixth case, "fluid" specifically had the meaning of water, which was translated into a word suggesting general liquid. In the seventh case, the word "response" corresponded to several words in Japanese, and the selection of words was not correct to represent the reduction of cancer cells.

Both "Mistranslation" and "Terminology" are the issue of word choice; however, there is a substantial difference in the two error types, as presented in these examples. Our typology design allowed distinguishing these two error types in a specific domain by fine-grained annotation.

6 Analysis of Annotation Results

We conducted an in-depth analysis of annotation results from four perspectives:

- Frequency and distribution of errors in current NMT systems (Section 6.1),
- Possible factors affecting error occurrence (Section 6.2),
- Co-occurrence of errors to reveal dependence among error types (Section 6.3), and
- Correlation with conventional automatic metrics for machine translation evaluation to investigate their powers of the test (Section 6.4).

6.1 Error Distribution

The rate of error occurrence was 1.98 per sentence, with a standard deviation of 2.07. The rate of error occurrence per source word was 0.09. This means that, on average, NMT outputs included approximately two errors within one sentence, although the high standard deviation suggested that the distribution of the presence of errors was somewhat dispersed. As shown in Figure 1, most of the sentences had errors of five or less (94.60%), and 572 sentences (23.06%) had no error.

Table 4 shows the distribution of errors by error types. Errors in terms of "Terminology" accounted for more than one-third. The second-largest proportion was "Mistranslation" (22.78%) followed by "Grammar" errors (20.38%).

6.2 Factors affecting to Error Occurrence

We investigated possible factors that may affect the occurrence of errors in NMT outputs. Namely, we investigated the effects of

Error type	Source	Target	Reference
Addition	Even former athletes who stop exercising do not retain measurable long-term benefits.	運動をやめた元スポーツ選手でさえ、長期的な (long-term) 長期的な (long-term) 利益を維持することはできない。	元運動選手であっても、運動をやめてしまえば、その効果を長期間維持することはできません。
Omission	Regular exercise can improve both of these qualities.	通常の運動は、これらの性質を改善することができる。	定期的な運動によってその両方 (both of) を向上させることができます。
Mistranslation	The primary end point was a composite of death, the need for dialysis, or a persistent increase of at least 50% from baseline in the serum creatinine level at 90 days.	主要なエンドポイントは、死亡、透析の必要性、または90日間 (for 90 days) の血清クレアチニンレベルのベースラインからの少なくとも50%の持続的な増加の複合物であった。	90日の時点 (at 90 days) における死亡、透析の必要性、血清クレアチニン値のベースラインから50%以上の上昇の持続の複合を主要評価項目とした。
Mistranslation	When men with BPH urinate, the bladder may not empty completely.	BPHの男性が排尿すると、膀胱が完全に空になることはありません (will not empty)。	前立腺肥大症の男性が排尿する場合、膀胱が完全に空にならないことがあります (may not empty)。
Grammar	Aging, estrogen deficiency, low vitamin D or calcium intake, and certain disorders can decrease the amounts of the components that maintain bone density and strength.	老化、エストロゲン欠乏、低ビタミンDまたはカルシウム摂取 (low vitamin D, and calcium intake)、およびある種の障害は、骨密度および強度を維持する成分の量を減少させる可能性がある。	加齢、エストロゲンの不足、ビタミンDやカルシウムの摂取不足 (low vitamin D or calcium intake)、およびある種の病気によって、骨密度や骨の強度を維持する成分の量が減少することがあります。
Terminology	Maintaining adequate levels of fluid and sodium helps prevent heat illnesses.	十分な量の液体 (liquid) とナトリウムを維持することは、熱病予防に役立ちます。	十分な水分 (water) およびナトリウム値を維持することが、熱中症の予防に役立つ。
Terminology	The rate of any complete or partial response to cabozantinib, vandetanib, and sunitinib was 37%, 18%, and 22%, respectively.	カボザンチブ、バンデタニブ、およびスニチニブに対する完全または部分応答 (answer) の割合は、それぞれ37%、18%および22%であった。	完全/部分奏効 (response) 率はCabozantinib 37%、Vandetanib 18%、およびSunitinib 22%であった。

Table 3: Examples of annotation results (Underlines indicate the errors with corresponding English translations in parentheses. Underlines and parentheses are for explanation and do not included in the actual annotation corpus.)

Subtype	Occurrence (%)	Mean per sentence (SD)
Addition	230 (4.68%)	0.09 (±0.40)
Omission	794 (16.16%)	0.32 (±0.73)
Mistranslation	1,119 (22.78%)	0.45 (±0.75)
Grammar	1,001 (20.38%)	0.40 (±0.74)
Terminology	1,768 (35.99%)	0.71 (±0.95)
Total	4,912 (100.00%)	1.98 (±2.07)

Table 4: Error occurrence based on the typology

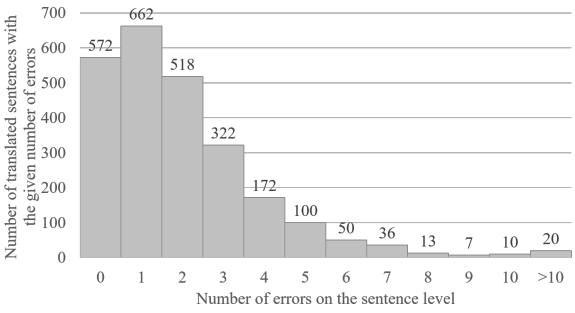

Figure 1: Distribution of errors in sentence ()

the length of source sentences, expertise level of source documents, and terminology.[6]

6.2.1 Length of Source Sentence

One of the most intuitive factors that affect the quality of NMT outputs is the length of the source sentence, i.e., longer sentences are more difficult to translate. As expected, source length was confirmed to have a high correlation with error occurrence. The correlation coefficients were $\rho = 0.65$ for the number of words in a sentence ($p < 0.0001$).

[6] These are dependent factors for each other, but we independently investigated their effects for simplicity.

6.2.2 Effect of Expertise Levels of Documents

We assumed that sentences from documents for experts were more challenging for NMT systems due to discrepancies in terminologies from those of the general domain. Among the sources of our corpora, two versions of MSD Manuals were about the same topics of medical information but distinguished by the levels of expertise: the Consumer Version was targeted at the general population, and the Professional Version was at health professionals. Source sentences of the Professional Version and the Consumer Version had 2,819 and 2,123 unique words, respectively, of which overlapped presence was limited to 984 words.

The difference in error occurrence was summarized in Table 5. Overall, translations of the Professional Version had a larger number of errors (1,108) than those of Consumer Version (770). Specifically, the errors of "Mistranslation," "Grammar," and "Terminology" were significantly more frequent on translations of Professional Version than on those of Consumer Version.[7] These results confirm our assumption that expertise levels of source documents negatively affect to the translation quality of current NMT systems.

6.2.3 Error Occurrence Dependent on Terms

Table 4 shows that the most common error types in NMT outputs are incorrect translations of terms, i.e., "Mistranslation" and "Terminology," which took up in total of 58.77% of errors. In this section, we further investigated what kind of words tend to cause these errors.

Table 6 ranks the most frequent words that were annotated as "Mistranslation" and "Terminology," respectively.[8] Frequent "Mistranslation" words included numbers and units ("days," and "months"), comparative words ("more," "less," and "versus"), and auxiliaries ("may"). In our analysis, these types of words more frequently produced incorrect translation than proper nouns, verbs, or other specific words in medicine. These words look simple but require different translations depending on co-occurring words and the context.

"Terminology" errors list different types of words from "Mistranslation." The high-ranked words such as "primary" and "response" are polysemous in the domain of medicine, which was failed to translate correctly by NMT systems.

6.3 Co-occurrence of Error Types

In this section, we investigated the interaction between error types to examine if some errors tend to lead to other types of errors. To determine the tendency of co-occurrence of the errors, we computed correlation coefficients of combinations of error types.

Table 7 shows combinations of error types whose correlation coefficients were larger than 0.3. The highest co-occurrence was observed in the combination of "Addition" and "Omission." Notably, in the total of 176 occurrences of "Addition" errors, 100 (56.82%) were accompanied by "Omission" errors. The errors of "Addition" and "Omission" were typically caused by over-generation and under-generation in NMT, respectively. This result revealed that over and under generations affect each other; over-generation of unnecessary phrases may lead to under generation of necessary phrases, and vice versa.

It is reasonable that "Addition" and "Omission" co-occur with "Grammar" errors, because the insertion of unnecessary words or deletion of necessary words may corrupt grammatical structures. The other way around is also possible, i.e., source sentences that an NMT system fails to capture correct grammatical structures are difficult to translate, which results in "Addition" and "Omission" errors.

The high co-occurrence of these errors suggests that the common problems of machine translation may mutually have causal correlations.

6.4 Correlation with Automatic Metrics

Finally, we investigated the correlation between annotated errors and BLEU scores as the most commonly used automatic evaluation metric. Specifically, we calculated a correlation coefficient between the number of errors in a sentence and sentence BLEU score. In addition, we also calculated the correlation with

[7] Although a significant difference was also confirmed on "Addition," we omit it due to their small numbers of occurrences.

[8] Stop words, such as short function words and punctuation marks, were filtered out from the ranking for brevity.

Subtype	Occurrence		p-value
	Consumer (Merck and Co., Inc., 2015a)	Professional (Merck and Co., Inc., 2015c)	
Addition	26	46	0.0300
Omission	142	168	0.1071
Mistranslation	225	265	0.0489
Grammar	102	224	< 0.0001
Terminology	275	405	0.0001
Total	770	1,108	

Table 5: Error occurrence by expertise levels of documents (Student t-test was used to calculate p-values)

Mistranslation		Terminology	
count	word	count	word
27	may	61	primary
15	more	33	response
14	days	33	common
12	less	28	survival
11	pneumonitis	28	outcome
10	rate	26	end
10	versus	22	point
9	common	19	fluid
9	therapy	18	active
9	months	17	benefit
9	active	17	therapy
8	falls	16	rate
7	medical	16	analysis
7	benefit	15	Secondary
7	drug	14	drug
7	ratio	14	overall
7	arms	14	ovarian
6	illness	14	studies
6	disease	13	outcomes
6	number	12	cancer

Table 6: Ranking of words with "Mistranslation" and "Terminology" errors

Error Combination	ρ	p value
Addition & Omission	0.43	< 0.0001
Omission & Grammar	0.35	< 0.0001
Addition & Grammar	0.31	< 0.0001

Table 7: Highly correlate error types ($\rho > 0.3$)

fairly simple metric, normalized Levenshtein distance between the translation outputs and reference translations as a baseline.

The correlation coefficient of error occurrence and sentence BLEU was $\rho = -0.18$ ($p < 0.0001$) while that of normalized Levenshtein distance was $\rho = 0.27$ ($p < 0.0001$). The sentence BLEU showed an even lower correlation than the normalized Levenshtein distance. This result indicates that sentence BLEU is not only ignorant of errors in translation output but also fails to evaluate the overall translation quality. Our annotation corpus contributes to design new automatic evaluation metrics that have the power to discriminate errors.

7 Discussion and Future Work

We performed the error analysis of NMT for the English and Japanese language pair in the medical domain, based on fine-grained and quality-controlled manual annotation.

In the analysis of detected 4,912 errors on 2,480 sentences, we found that the major errors in NMT were "Mistranslation" and "Terminology," rather than "Addition" and "Omission." The errors of "Addition" and "Omission" have been deemed typical in NMT as over-generation and under-generation, respectively; however, our results revealed that the semantic and terminology errors were more common in domain-specific technical documents. Interestingly, these errors were often observed in quantitative and polysemous words. This finding suggests future challenges in machine translation research targeting in the representation of numeric and multi-sense words.

We found more errors in documents for health-care professionals compared with those for the general public, specifically in terms of errors in "Grammar" and "Terminology." This finding encourages further research to improve the performance of NMT in documents that include sentences with complex syntax and highly-specialized technical terms.

The results of annotation will be published as a parallel corpus with detailed error labels, which is expected to be a valuable resource to improve NMT models, develop automatic evaluation metrics, and estimate qualities of machine translation. The limitations in current automatic evaluation metrics are partly attributable to insufficient understanding of the real performance of NMT systems. Furthermore, the dependence on the reference translation is problematic. The similarity to the reference does not necessarily represent the seman-

tic accordance of the translation to the source sentence. Natural language is characterized by its ambiguity, such as multiple meanings and contextual implications, and thus translation should not have the unique correct answer. While verbatim similarly to the reference enforces a strict constraint, it does not ensure the actual quality of translation. Better estimation of translation quality should incorporate features reflecting the actual quality of translation, such as semantic accuracy and linguistic fluency.

We believe our corpus contributes to research on evaluation or estimation models of NMT performance to overcome these limitations. Essentially, it is a valuable resource for assessing the domain-specificity of translation outputs. As future works, we will develop quality estimation models using the corpus to allow fine-grained and domain-specific evaluation. Also, we will extend the annotation corpus in other domains and language pairs.

Acknkowlegement

This work was supported by NTT communication science laboratories.

References

American Society of Clinical Oncology. 2018. Journal of Clinical Oncology (Japanese Version). http://usaco.jcoabstracts.jp/contents/.

American Society of Clinical Oncology. 2019. Journal of Clinical Oncology. http://ascopubs.org/journal/jco/.

J. Blatz et al. 2004. Confidence estimation for machine translation. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, pages 315–321.

F. Kepler et al. 2019. Unbabel's participation in the WMT19 translation quality estimation shared task. In Proceedings of the Fourth Conference on Machine Translation, pages 78–84.

F. Klubička, A. Toral, V. M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. Machine Translation, 32(3):195–215.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 388–395.

A. Lommel, H. Uszkoreit, A. Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. Tradumàtica, (12):0455–463.

Massachusetts Medical Society. 2019. The New England Journal of Medicine. https://www.nejm.org/.

Merck and Co., Inc. 2015a. MSD MANUAL Consumer Version. https://www.msdmanuals.com/home/.

Merck and Co., Inc. 2015b. MSD MANUAL Consumer Version (Japanese Version). https://www.msdmanuals.com/ja-jp/.

Merck and Co., Inc. 2015c. MSD MANUAL Professional Version. https://www.msdmanuals.com/professional/.

Merck and Co., Inc. 2015d. MSD MANUAL Professional Version (Japanese Version). https://www.msdmanuals.com/ja-jp/.

Nankodo Co.,Ltd. 2019. The New England Journal of Medicine (Japanese Version). https://www.nejm.jp/.

Pharmaceuticals and Medical Devices Agency. 2018. ICH guidelines (Japanese Version). https://www.pmda.go.jp/int-activities/int-harmony/ich/0070.html.

A. Rigouts Terryn et al. 2019. Pilot study on medical translations in lay language: Post-editing by language specialists, domain specialists or both? In Translating and the Computer 41. Editions Tradulex.

J. Singh. 2015. International conference on harmonization of technical requirements for registration of pharmaceuticals for human use. Journal of pharmacology & pharmacotherapeutics, 6(3):185.

L. Specia et al. 2017. Translation quality and productivity: A study on rich morphology languages. In Proceedings of Machine Translation Summit XVI, pages 55–71.

L. Specia et al. 2018. Findings of the wmt 2018 shared task on quality estimation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 689–709.

A. Tezcan, V. Hoste, L. Macken. 2017. Scate taxonomy and corpus of machine translation errors. In Trends in e-tools and resources for translators and interpreters, pages 219–244.

Z. Tu et al. 2016. Modeling coverage for neural machine translation. arXiv preprint arXiv:1601.04811.

L. Van Brussel, A. Tezcan, L. Macken. 2018. A fine-grained error analysis of nmt, smt and rbmt output for english-to-dutch. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

R. Wang et al. 2018. Sentence selection and weighting for neural machine translation domain adaptation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10):1727–1741.

Y. Wu et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.