An Empirical Study on the Robustness of Massively Multilingual Neural Machine Translation

Supryadi, Leiyu Pan, Deyi Xiong*
College of Intelligence and Computing, Tianjin University
Tianjin, China
{supryadi, lypan, dyxiong}@tju.edu.cn

Abstract

Massively multilingual neural machine translation (MMNMT) has been proven to enhance the translation quality of low-resource languages. In this paper, we empirically investigate the translation robustness of Indonesian-Chinese translation in the face of various naturally occurring noise. To assess this, we create a robustness evaluation benchmark dataset for Indonesian-Chinese translation. This dataset is automatically translated into Chinese using four NLLB-200 models of different sizes. We conduct both automatic and human evaluations. Our in-depth analysis reveal the correlations between translation error types and the types of noise present, how these correlations change across different model sizes, and the relationships between automatic evaluation indicators and human evaluation indicators. The dataset is publicly available at `https://github.com/tjunlp-lab/ID-ZH-MTRobustEval`.

Keywords: Multilingual Neural Machine Translation, Robustness, Evaluation

1. Introduction

Recent years have witnessed that neural machine translation (NMT) achieves a remarkable progress in both high- and low-resource language translation. For the former aspect, translation quality is substantially improved for many high-resource language pairs (e.g., Chinese-English, French-English) over the years, which has been tracked by yearly WMT evaluation (Bojar et al., 2018). Human parity has even been reached for some language pairs in terms of certain evaluation protocols (Hassan et al., 2018; Barrault et al., 2019). For the latter aspect, to improve translation quality of low-resource languages, massively multilingual neural machine translation (MMNMT) has been explored with growing interest, which enables knowledge transfer from high-resource languages to low-resource languages (Aharoni et al., 2019; Fan et al., 2021; Costa-jussà et al., 2022).

Conversely, NMT still faces challenges related to robustness, particularly in handling noise (Belinkov and Bisk, 2018) and adapting to domain shifts (Lai et al., 2022). In this study, we aim to delve into the translation robustness of Indonesian-Chinese within the context of massively multilingual NMT. Our specific objectives include understanding: 1) the patterns of the relationship between translation error types and noise types, and 2) how these patterns change across various MMNMT model sizes, ranging from models with millions to billions of parameters. Such an investigation holds significant importance in advancing our understanding of the robustness of Indonesian-Chinese translation, which remains an underexplored area, and in the development of MMNMT models.

To empirically study these patterns and relations, we use the open-sourced NLLB-200 (Costa-jussà et al., 2022) models as our MMNMT models. We curate an Indonesian-to-Chinese translation robustness evaluation dataset that consists of 1001 sentence pairs. Both languages are among the top-20 most spoken languages in the world but the parallel resources for them are very limited. We crawl noisy Indonesian sentences from social medias and manually translate them into Chinese with the collaboration between source language local speaker and the expert of the target language.

We manually identify noises in the source language and categorize them into 10 groups. These noisy source sentences are then automatically translated into Chinese with four NLLB-200 models of different sizes, where translation errors in translated target sentences are detected and classified into 10 categories. In addition to automatic evaluation of translation results with BLEU (Papineni et al., 2002) and CHRF++ (Popović, 2017), we also conduct human evaluation with multidimensional quality metric (MQM[1]).

The contributions of our work are as follows:

- We empirically evaluate the robustness for Indonesian-Chinese translation. To the best of our knowledge, this is the first attempt to study the robustness of Indonesian-Chinese translation based on MMNMT models.

- We curate a new noisy parallel dataset on Indonesian-Chinese translation for such evaluation.

*Corresponding author.

[1] `www.themqm.org`

LREC-COLING 2024, pages 1086–1097
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

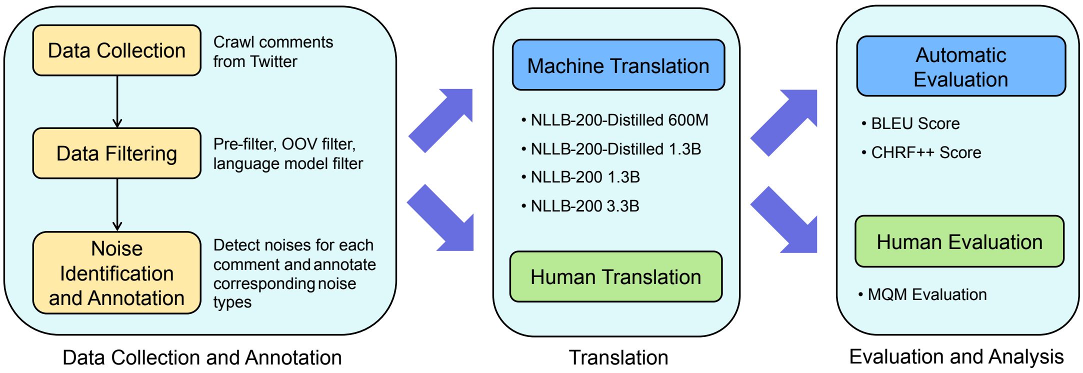
Data Collection
Crawl comments from Twitter
Data Filtering
Pre-filter, OOV filter, language model filter
Noise Identification and Annotation
Detect noises for each comment and annotate corresponding noise types
Data Collection and Annotation
Machine Translation
• NLLB-200-Distilled 600M
• NLLB-200-Distilled 1.3B
• NLLB-200 1.3B
• NLLB-200 3.3B
Human Translation
Translation
Automatic Evaluation
• BLEU Score
• CHRF++ Score
Human Evaluation
• MQM Evaluation
Evaluation and Analysis

Figure 1: Robustness evaluation and analysis protocol.

- We manually identify noise types in the dataset and translation error types in translations generated by the NLLB-200 models, study the relation patterns of them and examine the changes of these patterns across different model sizes.

2. Related Work

NMT Robustness and Evaluation Robustness is of paramount importance for neural machine translation (NMT), especially when NMT systems are deployed in real-world applications. A wide variety of efforts have been dedicated to enhancing the robustness of NMT. Among them, black-box methods are widely explored (Pinnis et al., 2017; Karpukhin et al., 2019; Liu et al., 2019; Wallace et al., 2020; Qin et al., 2021; Zhang et al., 2021). Alternatively, white-box methods, employing gradient-based approaches, have also been proposed (Cheng et al., 2019). Moreover, empirical evidence suggests that attacking NMT from the source side yields greater effectiveness (Zeng and Xiong, 2021). These methods usually employ synthetic noise to improve robustness.

In the context of robustness towards natural noise, the MTNT dataset (Michel and Neubig, 2018) is designed, originating from noisy data collected from Reddit[2] comments. This dataset comprises three different languages: English, French, and Japanese. In a similar vein, for the assessment of French-English translation robustness, noisy data have also been gathered from restaurant reviews (Berard et al., 2019). Additionally, for the evaluation of Chinese-English translation robustness, a dialogue dataset has been created as the natural noise data (Wang et al., 2021).

Partially inspired by Michel and Neubig (2018), we have curated a novel robustness evaluation dataset. However, our dataset differs significantly from Michel and Neubig (2018) in three aspects. Firstly, our primary focus lies in assessing the robustness of Indonesian-Chinese translation from two geographically distant languages. Secondly, our noisy data is derived from Twitter comments rather than Reddit, encompassing a broader spectrum of topics. Thirdly, we have manually identified and annotated different noise types for each sentence pair, enabling a more targeted evaluation of noise-specific robustness.

Multilingual NMT Multilingual neural machine translation (MNMT) has garnered growing interest in recent years owing to its capacity to facilitate the deployment of NMT systems supporting multiple languages, knowledge transfer between languages (Sun and Xiong, 2022), and zero-shot translation capabilities, among others (Xu et al., 2021; Li et al., 2023). To enable knowledge transfer across an extensive array of languages, including 100 or more languages, research has delved into massively multilingual neural machine translation (MMNMT) (Johnson et al., 2017; Jin and Xiong, 2022). This exploration has evolved from English-centric models (Aharoni et al., 2019) to models extending beyond English-centric approaches, such as M2M-100 (Fan et al., 2021).

Among those non-English-centric models, NLLB-200 (Costa-jussà et al., 2022) has recently been open-sourced, which encompasses 200 languages and 40,000 translation directions, supported by a model with up to 54 billion parameters trained on a huge amount of natural and synthesized data. In this study, we employ NLLB-200 models to assess the robustness of MMNMT on non-English languages using our curated dataset.

Multilingual NMT Robustness The robustness of multilingual NMT also been evaluated recently (Pan et al., 2023). A variety of noises at the character-, word-, and multiple levels have been explored for the study of multilingual NMT robust-

[2] www.reddit.com

ness. It has been observed that the robustness of multilingual NMT can be transferred across languages. In contrast to previous study, this research specifically focuses evaluating the robustness of MMNMT towards naturally occurring noise, which is categorized into 10 distinct types.

3. Robustness and Evaluation Protocol

We propose a general protocol for evaluating and analyzing MMNMT robustness towards naturally occurring noises, which is model- and language-independent. As illustrated in Figure 1, the protocol consists of three main stages.

1) **Data Collection and Annotation**: In this initial stage, we commence by identifying suitable sources for collecting noisy data. Once the sources are determined, data is extracted from these sources. We employ automatic noise detection methods to filter the extracted data, retaining only the noisy portions. In our study, this extraction and detection process yields a high-quality monolingual Indonesian corpus that incorporates naturally occurring noise. Each sentence in this corpus is then labeled with its associated noise category. It is worth noting that each sentence may be annotated with multiple noise categories.

2) **Translation**: The collected source corpus is then translated into the target language by human translators, adhering to a noise translation convention to ensure consistency in the translation of noisy fragments throughout the entire corpus. These manual translations serve as reference translations for both automatic evaluation and manual analysis. To assess and analyze the robustness of specific MMNMT models, the source corpus is also automatically translated into the target language by these MMNMT models.

3) **Evaluation and Analysis**: In this stage, we carry out both automatic and human evaluations. We manually identify translation errors and categorize them according to multidimensional quality metrics (MQM) (specifically level-1 error types). With annotated noise types and translation error types, we can conduct a thorough and comprehensive analysis of the robustness issues observed in MMNMT models.

We will detail the data collection and annotation procedure in Section 4, translation in Section 5 and evaluation in Section 6. In-depth analysis results are presented in Section 7.

4. Data Collection and Annotation

4.1. Data Collection

We collected raw social media comments from Twitter. To obtain these comments, we utilized Tweepy[3], a Python library for accessing the Twitter API. Given our focus on the Indonesian language, the comments are crawled using popular Twitter accounts from Indonesia as keywords. The collection period for these comments spanned one week, from 13 December 2022 to 20 December 2022. The final dataset contains a total of 25,973 comments.

4.2. Data Filtering

After the collection of these comments, we employed filtering methods to detect noisy comments, following the approach outlined by Michel and Neubig (2018) in the MTNT dataset. We utilized three filtering methods for this purpose.

Pre-filter We performed a pre-filtering process on the collected raw data in three steps, with the aim of retaining only naturally occurring noises:

1) Removing comments containing URLs.

2) Removing comments from users where their usernames contain “bot” or “AutoModerator”.

3) Removing comments written in other languages. We use Python library Langid.py[4] to detect non-Indonesian languages.

OOV Filter For robustness evaluation, we aimed to make the corpus as noisy as possible, considering out-of-vocabulary (OOV) words as a form of noise. To introduce unknown words and add noise to the sentences, we created a dictionary using a contrast corpus. Our contrast corpus comprises the Indonesian section of WMT20news-commentary-v15[5] and OpenSubtitles[6]. Using the fairseq tool (Ott et al., 2019), we generate a dictionary containing 5,000 words. Then, we only keep those comments that contain at least one OOV word.

Language Model Filter In the final step, we employed an n-gram language model to further identify noisy comments. We tokenized both the contrast corpus and the collected comments using Byte-Pair Encoding (BPE) with SentencePiece (Kudo and Richardson, 2018). We then trained a

[3] https://github.com/tweepy

[4] https://github.com/saffsd/langid.py

[5] https://www.statmt.org/wmt20

[6] https://www.opensubtitles.com

Indonesia Text	Spelling/Typo	Grammar	Spoken Language	Slang	Proper Noun	Dialect	Code Switching	Jargon	Emojis	Slurs
Makasih kang sial. Berkat lu dukung prancis argentina jadi juara. Sekali lagi terima kasih sudah bikin prancis sial			V			V	V		V	
alah ribet amat urusan ucap natal ga ucap nata, noh negara lain udah mikirin hidup di mars.	V	V	V	V	V					
Gara2 ngomen denny Akun.ku dihanguskan njir		V	V					V		V

Table 1: Noise identification and annotation.

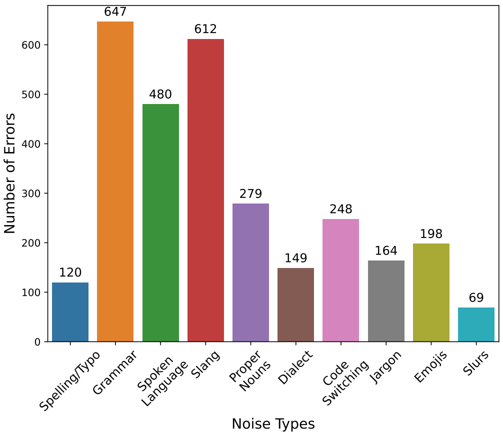

Figure 2: Statistics of different noise types in the curated dataset.

5-gram Kneser-Ney smoothed language model on the segmented contrast corpus. This trained model is used to calculate language model scores for the comments, normalized by sentence length. We selected comments within a specific score interval, specifically the range between the first and third quartiles of normalized language model scores in the comment corpus. However, we ensured that the normalized language model score of a retained comment was smaller than the third quartile of normalized language model scores in the contrast corpus. This approach strikes a balance, ensuring that each kept comment contains a certain amount of noise without being overly noisy, which differs from the method used in MTNT.

After applying the three filters to the collected raw data, we retained 1,001 comments in the final dataset. The average sentence length of these retained comments is 10 words, with a standard deviation of 6.6. The shortest comment consists of 1 word, while the longest comment comprises 48 words.

4.3. Noise Identification and Annotation

We use a noise taxonomy similar to that used in MTNT, which consists of:

- **Spelling/typographical errors**: Comments contain incorrectly spelled or typed words.
- **Grammatical errors**: Comments are not grammatically written.
- **Spoken language**: Comments are written in the style of spoken language.
- **Internet slang**: Comments contain trending words in internet/social media.
- **Proper nouns**: Proper nouns, e.g., entities of place, person, are incorrectly written.
- **Dialects**: Comments contain Indonesia dialects, e.g., Javanese, Acehnese, and Balinese.
- **Code switching**: Comments contain more than one language.
- **Jargon**: Comments include specific words used in certain areas of life (environment).
- **Emojis**: Comments contain emojis for feeling expression.
- **Slurs**: Improper words are used to insult people.

In the corpus, we detected each instance of noise and classified them into their respective noise types, as previously described. It is worth emphasizing that a sentence may contain multiple types of noise, and we annotate each of these noise types for the sentence accordingly.

The example of our noise identifications and annotations is presented in Table 1. In the first sentence, “kan” is a Sundanese dialect with the meaning “brother”, and “lu” is a Hokkien dialect meaning “you”. These two words indicate the presence of code switching in the sentence. Additionally, the word “prancis” is a proper noun referring to “France”, which should be capitalized as “Prancis”. Towards the end of the sentence, the spoken word “bikin” should be replaced with the written word “buat”, and an emoticon is present at the end of the sentence.

In the second sentence, the first word “alah” is an interjection expressing “complaining”. There is a missing conjunction “and” between “ucap natal” and “ga ucap natal”. “Ga” is a slang term that translates to “no” in English. The word “nata” appears to be a typographical error and should be corrected to “Natal”, which means “Christmas”. Similarly, “mikirin” is a spoken language form and should be written as “memikirkan”, which means “thinking”. Finally, “Mars” is a proper noun referring to the name of a planet.

Error Type		Description
Terminology		Inconsistency and accuracy issues of the terminology.
Accuracy	Mistranslation	Target content that does not accurately represent the source content.
	Omission	The target content is missing from the translation that is present in the source.
	Addition	Target content that includes content not present in the source.
	Untranslated	The text in source content is left untranslated in the target content.
	Hallucination	The translation is very different or irrevelant with the source.
Fluency	Grammar	The translation result violates the grammatical rules of the target language.
	Punctuation	Incorrect punctuation for the locale or style.
Local Convention		The translation violates locale-specific content or formatting requirements.
Audience Appropriateness		The use of content in the translation that is invalid or inappropriate for the target audience.

Table 2: MQM hierarchy.

Severity Level	Description
Critical	The errors that significantly affect translation usability, understandability, and meaning.
Major	Errors that would impact usability or understandability of the translation.
Minor	Errors that would not impact the usability or understandability of the translation.

Table 3: MQM severity levels.

In the last sentence, “ngomen” is a spoken language form that should be written as “mengomentari”, which means “to give comments”. The use of “.k” is a grammar error since the sentence is not finished, but a full stop is placed in the middle. Furthermore, it is important to note that “njir” is a derogatory slang term in the Indonesian language that dehumanizes individuals by comparing them to dogs.

The statistics for annotated noise types are presented in Figure 2. The most prevalent noise type is grammatical noise, which is observed in 647 comments. Additionally, the spoken language and slang noise types are also prominently represented. This observation aligns well with our expectations for social media texts.

5. Translation

The annotated corpus is then translated into Chinese manually, creating a parallel corpus that acts as a benchmark testbed for evaluating robustness. We enlisted language experts proficient in both the source and target languages for this task. Through collaboration with experts in both languages, the sentences’ meanings are conveyed with greater accuracy, making the translations more comprehensible to the reader. These translations were reviewed and proofread by professional translator to ensure translation consistency, particularly in noisy parts of the corpus. To curate a high-quality parallel corpus, it is important to establish guidelines ensuring consistency in the translation results. Our guidelines for the parallel corpus translation include:

- Translating punctuation according to the target language convention and normalizing its usage.
- Translating idioms to convey their meaning and ensure reader comprehension.
- Correcting grammar errors in the source sentences during translation, maintaining proper grammar.
- Standardizing the translation of proper nouns (names of people, places, organizations, products) across the entire corpus.
- Preserving emojis in the translation.
- Standardizing the translation of slang throughout the corpus.

For machine translation, we utilized NLLB-200, a recently released MMNMT model. We employed four variants of the NLLB-200 model for automatically translating the collected dataset: NLLB-200-Distilled 600M, NLLB-200-Distilled 1.3B, NLLB-200 1.3B, and NLLB-200 3.3B. The results of machine translation will be evaluated and analyzed.

6. Evaluation

An evaluation is conducted to compare the translation results of NLLB across different model sizes. By performing human and machine translations on the corpus, we conducted both automatic and human evaluations to assess the robustness of the NLLB-200 model against naturally occurring noise.

6.1. Evaluation Settings

For each NLLB model translation results, we conduct the automatic and human evaluation. The source and reference used for the evaluation is from the curated Indonesian-Chinese dataset.

For automatic evaluation, we use automatic metrics of BLEU and CHRF++, which collectively measure both word- and character-level translation

Evaluation	NLLB-200-Distilled 600M	NLLB-200-Distilled 1.3B	NLLB-200 1.3B	NLLB-200 3.3B
BLEU	11.43	10.96	12.58	**14.04**
CHRF++	9.56	9.34	10.33	**11.23**
MQM	2.21	5.54	10.70	**12.17**

Table 4: BLEU, CHRF++, and MQM scores of translation results yielded by different models on the dataset.

quality. We use SacreBLEU tool for calculating the BLEU and CHRF++ score (Post, 2018).

Additionally, human evaluation is conducted to enhance the evaluation results. To achieve this, we utilize various types of translation errors from the multidimensional quality metric (MQM) framework, which are categorized into five groups: *terminology*, *accuracy*, *fluency*, *local convention*, and *audience appropriateness*. Accuracy encompasses translation errors such as *mistranslation*, *omission*, *addition*, *untranslated*, and *hallucination*. Fluency covers translation errors related to *grammar* and *punctuation*. In NMT, there is a possibility that the NMT system may produce strange or irrelevant translations. Therefore, in our experiment, we introduce a new error type called "hallucination", which is not included in the MQM framework. The detailed hierarchy of the error types is presented in Table 2.

Among the ten previously mentioned error types (ET), each one is initially given a standard error weight of 1. However, in the case of hallucination, we assign a weight of 3. This choice is grounded in the recognition that hallucination represents an exceptionally critical error, as it involves a substantial departure in meaning from the source sentence.

Each error type (ET) has three severity levels: *minor*, *major*, and *critical*, with multiplier scores of 1, 5, and 10, respectively. The detail explanation of each severity level is shown in Table 3. We manually identify translation errors for each target translation and annotate the corresponding translation error type and severity level. Additionally, we permit the annotation of multiple translation error types for each translation if different translation errors are found in the target translation.

Following the annotation process, we proceed to calculate the Overall Quality Score (OQS) using the MQM framework. To begin, we calculate the Error Type Penalty Total (ETPT) for each error type, as defined in Equation 1. Subsequently, the OQS is derived by evaluating the relationship between ETPT and the Evaluation Word Count (EWC), as described in Equation 2. The EWC denotes the total count of words present in the source language corpus.

$$\begin{aligned}\mathrm{ETPT} = (&\mathrm{ET}_{\mathrm{minor}} + \mathrm{ET}_{\mathrm{major}} \times 5 + \\ &\mathrm{ET}_{\mathrm{critical}} \times 10) \times \mathrm{ET}_{\mathrm{weight}}\end{aligned} \tag{1}$$

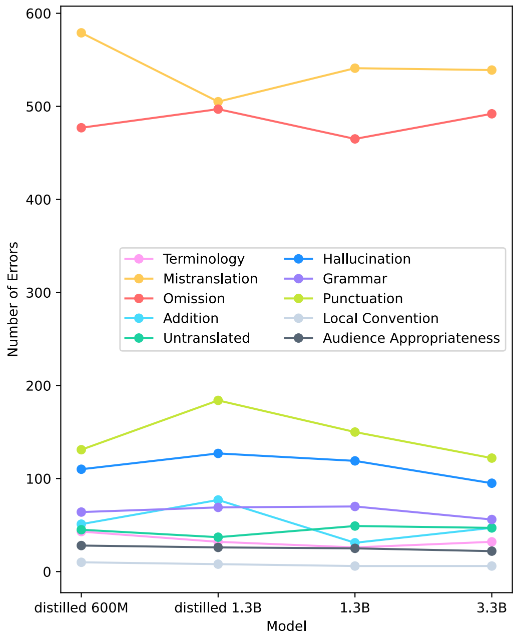

Figure 3: The change of translation error types with the increment of model parameters.

$$\mathrm{OQS} = \left(1 - \frac{\sum_i (\mathrm{ETPT}_i)}{\mathrm{EWC}}\right) \times 100 \tag{2}$$

6.2. Evaluation Results

First, we conducted an automated evaluation of Indonesian to Chinese translations produced by various models using the BLEU and CHRF++ metrics. The results are represented in Table 4. Additionally, the results of human evaluation using MQM are also included in Table 4. It can be observed that the performance of the models improves as their size increases.

In the human evaluation, Table 5 shows the occurrences of translation error types corresponding to different noise types across various models. Based on these findings, we conducted further analysis of the evaluation results in Section 7.

Model Size	Error Type	Spell/Typo Error	Grammar	Spoken	Slang	Proper	Dialect	Code switch	Jargon	Emojis	Slurs
NLLB-200-Distilled 600M	**Terminology**	6	26	23	16	11	5	5	**29**	2	1
	Mistranslation	80	**396**	292	369	172	83	147	85	108	47
	Omission	64	**335**	259	317	146	83	137	96	102	34
	Addition	6	**31**	24	29	10	5	7	8	6	3
	Untranslated	7	**32**	20	28	28	5	12	8	8	1
	Hallucination	13	63	58	**81**	23	28	37	11	27	13
	Grammar	3	**38**	16	32	15	9	15	8	15	3
	Punctuation	18	89	70	**101**	39	26	39	12	30	14
	Local Convention	3	**5**	2	4	3	0	0	1	1	0
	Audience Appropriateness	1	13	11	**19**	8	3	6	4	3	1
NLLB-200-Distilled 1.3B	**Terminology**	6	18	21	10	1	1	1	**26**	1	0
	Mistranslation	64	**339**	255	315	154	82	135	92	101	39
	Omission	66	**349**	256	343	152	82	138	92	119	30
	Addition	10	47	41	**49**	28	13	17	10	10	5
	Untranslated	8	23	14	**24**	19	7	10	9	5	3
	Hallucination	14	82	67	**85**	29	22	37	11	25	17
	Grammar	6	**42**	25	38	24	10	13	8	12	5
	Punctuation	29	**138**	93	131	52	32	50	30	46	15
	Local Convention	1	**3**	2	3	2	0	0	3	0	0
	Audience Appropriateness	1	11	9	**15**	6	2	4	3	4	0
NLLB-200 1.3B	**Terminology**	5	14	17	7	1	1	1	**22**	1	0
	Mistranslation	76	**361**	270	346	178	84	148	92	104	39
	Omission	69	**330**	240	295	137	82	138	92	102	32
	Addition	4	16	11	**19**	7	6	6	5	3	2
	Untranslated	8	**35**	23	32	30	6	15	10	8	4
	Hallucination	12	64	65	**85**	21	23	32	14	25	13
	Grammar	1	**46**	26	37	21	6	9	12	18	2
	Punctuation	25	104	73	**105**	32	34	54	22	35	20
	Local Convention	**2**	**2**	1	**2**	**2**	0	0	0	0	0
	Audience Appropriateness	1	11	9	**15**	6	1	3	4	4	1
NLLB-200 3.3B	**Terminology**	6	19	20	10	1	1	1	**27**	1	0
	Mistranslation	76	**352**	270	344	161	89	155	89	110	43
	Omission	70	**347**	251	330	153	81	135	91	103	33
	Addition	6	20	24	**24**	12	4	9	4	6	2
	Untranslated	11	34	23	**34**	26	8	15	12	9	1
	Hallucination	5	60	48	**67**	22	14	23	8	16	12
	Grammar	7	**31**	18	22	14	3	7	11	13	4
	Punctuation	20	**88**	59	82	41	24	36	15	33	12
	Local Convention	0	**2**	**2**	**2**	**2**	0	0	**2**	0	0
	Audience Appropriateness	0	8	9	**14**	4	2	3	2	3	1

Table 5: The number of translation error types corresponding to different models and noise types on the dataset.

7. Analysis

7.1. Effect of Model Size

Figure 3 illustrates the variation in translation error types as the number of model parameters increases. We conducted analysis for each model size in comparison to the subsequent larger model size.

In the comparison between distilled 600M and distilled 1.3B model, the number of *mistranslation* errors in the distilled 1.3B model is considerably lower than those in the distilled 600M model. But the number of *omission*, *addition*, and *punctuation* translation error types in the distilled 1.3B model is notably higher compared to the distilled 600M model. We speculate that the model is becoming more proficient in addressing mistranslation. However, it may encounter issues with yielding target translations incompletely or even producing an excessive translation.

Furthermore, when comparing the performance of the distilled 1.3B and 1.3B models, we observed that despite having the same amount of parameters, the 1.3B model exhibits more consistent performance across the 10 translation error types. On the other hand, the distilled 1.3B model demonstrates strong performance in terms of *mistranslation*, but relatively weaker performance in other translation errors. Hence, we conclude that model distillation may enhance the model's proficiency in specific areas while potentially compromising its capability in other areas. Conversely, a model trained directly without distillation may offer a more balanced performance across all areas.

As the model size expands to 3.3B, there is a reduction in the occurence of most error types, with the exception of *mistranslation*, which remains stable, and *addition*, which experiences a substantial increase. This suggests that the increase in model size may lead the MMNMT model to generate additional information.

7.2. Effect of Noise Types

In Table 5, the occurrence of translation error types based on different noise types is similar across various model sizes. With the exception of *addition*, *untranslated*, and *punctuation*, several models are affected by grammar or slang. However, the differences in counts are not significant. Therefore, we aggregate the results from the 4 models for further analysis.

Figure 4 shows a heatmap that assists in analyzing the occurrence of each translation error type based on the noise types present in the source sentences. The values in the heatmap represent the combined translation results from 4 models, which have been normalized according to each translation error type. The occurrence of *terminol-*

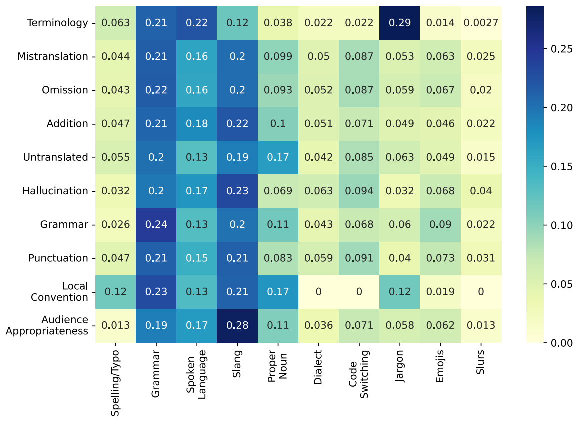

Figure 4: The heatmap of the occurences each of translation errors types according to the noise types.

	BLEU	CHRF++
MQM	0.8576	0.8741

Table 6: Pearson's correlation coefficient between automatically evaluated indicators and human evaluated indicators.

ogy errors primarily arises from existing jargon in the source sentences. Translation errors related to *accuracy*, *fluency*, and *local convention* primarily arise from slang and grammar noise within the source sentences. Furthermore, errors in *audience appropriateness* predominantly result from slang utilized in the source sentences.

7.3. Relationship between Automatic and Human Evaluation

It is evident that as the model size increases, both BLEU and CHRF++ scores exhibit a corresponding increase. Moreover, in average, they consistently align with human evaluation results, as measured by MQM. However, there is a slight divergence in the case of the NLLB-200-Distilled-1.3B model size, where there is a slight degradation in both BLEU and CHRF scores. To quantitatively evaluate the correlation between automatic and human assessment metrics, we calculated Pearson's correlation coefficient between the scores obtained from the automatic evaluation indicators and the number of translation errors identified by the human evaluation metrics. The results are presented in Table 6. Based on these findings, we can conclude that CHRF++ demonstrates a stronger correlation with human evaluation when compared to the BLEU score, signifying that CHRF++ serves as a more dependable automatic evaluation metric.

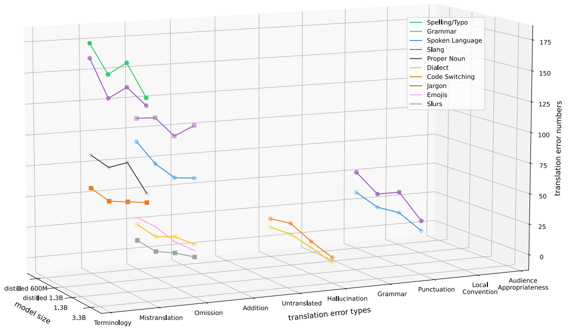

Figure 5: The changing trend of the number of translation errors along with the change of model parameters on short sentences. Dot represents the downward of the trends and square represents the upward of the trends.

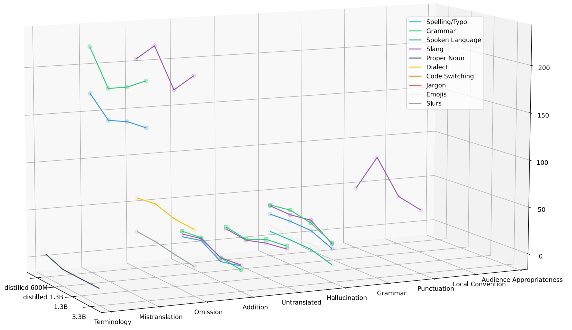

Figure 6: The changing trend of the number of translation errors along with the change of model parameters on long sentences. Dot represents the downward of the trends and square represents the upward of the trends.

7.4. Relationship between Sentence Length, Noise Types, and Model Sizes

When labelling the model translation results with translation error types, we found that there are several differences in the distribution of translation error types for short and long sentences. Thus, we analyzed short and long sentences separately and observed the correlations and differences. We first took the average length of sentences as the threshold value to differentiate long sentences from short sentences. Then we counted the number of translation errors corresponding to different noise types of different models, and fit the number of translation errors generated by different models under different noise types according to the linear regression method. If the primary term coefficient is less than -5 or greater than 0, it means that the number of translation errors has an obvious decreasing or increasing trend as the number of model parameters increases, and we will focus on these cases.

The results are shown in Figure 5 and Figure 6. The dots on the line represent a clear downward trend in the line, corresponding to a primary term coefficient less than -5. Conversely, the squares represent an upward trend in the line, corresponding to a primary term coefficient greater than 0. The transparency of both the dots and squares indicates the magnitude of these trends, with greater transparency reflecting stronger upward or downward trends.

First, we focus on the analysis of short sentences. In Figure 5, it can be observed that as the model size increases, a clear reduction in the number of translation errors related to grammar, slang, spoken language, and proper nouns noise types is evident. Furthermore, there is a noteworthy decline in the number of translation errors attributed to slang and spoken noise types for *punctuation*, indicating an enhancement in fluency. Nevertheless, concerning accuracy, a consistent upward trend is noticeable in the number of translation errors stemming from specific noise types. For instance, *omission* errors associated with slang, dialect, and slurs noise contribute to this ascending trend.

In the case of long sentences, a noticeable pattern emerges as the model size increases. We observed a substantial reduction in the number of translation errors associated with slang, grammar, and spoken language noise types, particularly in terms of accuracy. However, there was a tendency for an increase in *untranslated* translation errors. In terms of fluency, we observed a diminishing trend in *punctuation* errors. Upon comparing these findings, it becomes evident that longer sentences display improved performance with larger models when contrasted with shorter sentences.

8. Conclusions and Future Work

The robustness of NMT still poses challenges, especially towards natural occuring noises. Based on research findings, it can be concluded that the size of the model significantly impacts translation performance. In terms of noise types, it is evident that spelling and typographical errors can lead to inaccuracies, fluency issues, and translation errors related to terminology. Larger models perform better on longer sentences.

In the future, we aim to evaluate the robustness of low-resource languages using benchmark datasets. Additionally, given the emergence of large language models (LLMs), we plan to delve into evaluating their performance in translation tasks in comparison to traditional NMT models.

Limitations

Our experiments primarily rely on a curated Indonesian-Chinese parallel corpus crawled from Twitter comments with various types of noise. The dataset covers translations only from Indonesian to Chinese and serves as the evaluation benchmark. The dataset size is relatively small for training but is suitable for robustness evaluation. Due to limitation in computational resources, the NLLB-200 54B model is not used in this research.

Ethics Statement

Data Privacy The curated noisy parallel corpus of Indonesian-Chinese is openly available for research purposes. One concern regarding this data is that it is obtained by crawling Twitter comments, raising privacy concerns for Twitter users. In order to protect the privacy of Twitter users, we have removed user IDs and usernames from the dataset. In Twitter comments, it is common for users to tag other users by their usernames. To address this, we have systematically removed any text that includes the "@" prefix, thereby effectively eliminating tagged usernames.

Social Impact However, it is essential to note that the dataset is sourced from social media comments, where certain comments may not be appropriate for all audiences, including those containing hate speech or offensive language. In our experiments, we maintain these comments for the purpose of robustness evaluation, as they are also considered as a form of natural noise. It is crucial to take this aspect into account before utilizing our parallel corpus for other studies.

Acknowledgments

The present research was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01D43). We would like to appreciate the insightful comments provided by anonymous reviewers. We extend our gratitude to Ms. Yu for her assistance in proofreading the translation of the parallel corpus. Additionally, we gratitudely acknowledge the support of the Chinese Government Scholarship (CGS).

9. Bibliographical References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 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, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Renren Jin and Deyi Xiong. 2022. Informative language representation learning for massively multilingual neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5158–5174, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Wen Lai, Jindřich Libovický, and Alexander Fraser. 2022. Improving both domain robustness and domain adaptability in machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5191–5204, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shangjie Li, Xiangpeng Wei, Shaolin Zhu, Jun Xie, Baosong Yang, and Deyi Xiong. 2023. MMNMT: Modularizing multilingual neural machine translation with flexibly assembled MoE and dense blocks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4978–4990, Singapore. Association for Computational Linguistics.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Leiyu Pan, Supryadi, and Deyi Xiong. 2023. Is robustness transferable across languages in multilingual neural machine translation? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14114–14125, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data. In *Text, Speech, and Dialogue*, pages 237–245, Cham. Springer International Publishing.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Wenjie Qin, Xiang Li, Yuhui Sun, Deyi Xiong, Jianwei Cui, and Bin Wang. 2021. Modeling homophone noise for robust neural machine translation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7533–7537.

Haoran Sun and Deyi Xiong. 2022. Language branch gated multilingual neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5046–5053, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Eric Wallace, Mitchell Stern, and Dawn Song. 2020. Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5531–5546, Online. Association for Computational Linguistics.

Tao Wang, Chengqi Zhao, Mingxuan Wang, Lei Li, and Deyi Xiong. 2021. Autocorrect in the process of translation — multi-task learning improves dialogue machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 105–112, Online. Association for Computational Linguistics.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, and Deyi Xiong. 2021. Modeling task-aware MIMO cardinality for efficient multilingual neural machine translation. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 361–367, Online. Association for Computational Linguistics.

Zhiyuan Zeng and Deyi Xiong. 2021. An empirical study on adversarial attack on NMT: Languages and positions matter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 454–460, Online. Association for Computational Linguistics.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977, Online. Association for Computational Linguistics.