MQM-Chat: Multidimensional Quality Metrics for Chat Translation

Yunmeng Li[1] **Jun Suzuki**[1,2] **Makoto Morishita**[3,1] **Kaori Abe**[4,1,2] **Kentaro Inui**[5,1,2]
[1]Tohoku University [2]RIKEN
[3]Future Corporation [4]Machine Learning Solutions Inc. [5]MBZUAI
li.yunmeng.r1@dc.tohoku.ac.jp

Abstract

The complexities of chats, such as the stylized contents specific to source segments and dialogue consistency, pose significant challenges for machine translation. Recognizing the need for a precise evaluation metric to address the issues associated with chat translation, this study introduces Multidimensional Quality Metrics for Chat Translation (MQM-Chat), which encompasses seven error types, including three specifically designed for chat translations: ambiguity and disambiguation, buzzword or loanword issues, and dialogue inconsistency. In this study, human annotations were applied to the translations of chat data generated by five translation models. Based on the error distribution of MQM-Chat and the performance of relabeling errors into chat-specific types, we concluded that MQM-Chat effectively classified the errors while highlighting chat-specific issues explicitly. The results demonstrate that MQM-Chat can qualify both the lexical accuracy and semantical accuracy of translation models in chat translation tasks.

1 Introduction

Recent developments in neural machine translation (NMT) (Bahdanau et al., 2014; Patil and Davies, 2014; Medvedev, 2016; Gehring et al., 2017) have demonstrated notable improvements in the performance of machine translation systems, especially in tasks involving the translation of formatted documents such as news articles and academic papers (Maruf and Haffari, 2018; Barrault et al., 2019, 2020; Nakazawa et al., 2019; Ma et al., 2020). However, current methods continue to face considerable challenges when translating chats (Tiedemann and Scherrer, 2017; Maruf et al., 2018; Farajian et al., 2020) due to their high degrees of ambiguity and stylized contents, including sentiments, personalities, and cultural nuances (Uthus and Aha, 2013; Läubli et al., 2018; Toral et al., 2018; Farajian et al., 2020) To enhance the performance of chat translation, it is important to understand the qualities and limitations of existing translation models in handling chats. Traditional automatic evaluation metrics, such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022a,b), primarily focus on accuracy; however, they fail to capture the meanings and the stylized contents, especially when evaluating chats. For example, as shown in Table 1, translation models generate errors because of source-side typographical errors (typos), Internet slang, and the omission of subjects in the flow of chats. When evaluating the chat translation quality, we need to focus on nuances, stylized content specific to the source segments, and dialogue consistency in addition to grammatical and lexical accuracy. Thus, a refined error categorization framework that can assess semantic accuracy while preserving the chat-specific nuances and content is better suited for chat translation tasks (Gehman et al., 2020).

In this paper, we propose the Multidimensional Quality Metrics for Chat Translation (MQM-Chat) to address the challenges of evaluating chat translations. Building on the existing Multidimensional Quality Metrics (MQM) framework[1] (Burchardt, 2013; Mariana, 2014; Freitag et al., 2021a), MQM-Chat incorporates seven error types: *Mistranslation*, *Omission or Addition*, *Terminology or Proper Noun Errors*, *Unnatural Style*, *Ambiguity and Disambiguation*, *Buzzword or Loanword Issues*, and *Dialogue Inconsistency*. The latter three were specifically designed to handle the chat nuances.

MQM-Chat was applied to evaluate the chat translation capabilities of five models: the large language models (LLMs) GPT-4 (Achiam et al., 2023) and LLaMA3 (Touvron et al., 2024), the commercial translation model DeepL[2], the multilingual model by Facebook at WMT21 (Tran et al.,

[1] https://themqm.org/
[2] https://www.deepl.com/translator

Proceedings of the 31st International Conference on Computational Linguistics, pages 3283–3299
January 19–24, 2025.

Source (zh, ja)	**Possible Good Translation (en)**	**Bad Translation (en) by MT Model**
	Ambiguity and Disambiguation	
队啊！你应该试试！	Yaas! You should try!	Team ah! You should try!
知ってｒ？昨日、ヘレンとあったよ	u know waht, I saw Helen yesterday	You know what, I saw Helen yesterday
	Buzzword or Loanword Issues	
鼠的，真的累死了	Yaap, I'm really tired	Rat's, I'm really exhausted
草wwwwww	lol	grass
	Dialogue Inconsistency	
你觉得我能赢吗? - 没事儿，肯定能赢啊!	Do you think I can win? - It's okay. I'm sure you'll win!	Do you think I can win? - It's okay. I'm sure they'll win!
まどかは昨日買い物に行ったよ - 行った？聞いてないよ！	Madoka went shopping yesterday. - She went? I didn't hear about it!	Madoka went shopping yesterday. - You went? I didn't hear about it!

Table 1: Examples of MQM-Chat's chat-specific errors, including *Ambiguity and Disambiguation*, *Buzzword or Loanword Issues*, and *Dialogue Inconsistency*.

2021), and the bilingual model by team SKIM at WMT23 (Kudo et al., 2023). We translated Chinese (zh) and Japanese (ja) chat data into English (en) using these models and assigned human annotations. MQM-Chat helped highlight the issues and qualified the strengths and weaknesses of the five models crossing language pairs.

To verify the effectiveness of MQM-Chat, we compare it with the standard MQM framework. Standard MQM human annotations were assigned to sampled data, and the differences between the two approaches were analyzed. Our findings demonstrate that MQM-Chat provides fine-grained classification, recognizing a significant portion of errors in standard MQM into other labels, with approximately 30% of these as chat-specific issues such as *Ambiguity and Disambiguation*, *Buzzword or Loanword Issues*, and *Dialogue Inconsistency*.

We have attempted to implement automatic annotation using MQM-Chat with few-shot learning. The results indicated that the auto annotations obtained by MQM-Chat agree with the overall system performance annotated by human annotators but are not as accurate as human annotations.

In summary, this study contributes to the chat translation field by proposing the MQM-Chat evaluation metric. Five state-of-the-art translation models were evaluated with MQM-Chat when handling chat content. The experiments helped construct annotated zh⇒en and ja⇒en chat translation data. The contributions of this study are expected to enhance the understanding of chat translation and provide valuable resources for future advancements in the field.

2 Related Work

2.1 Translation Evaluation Metrics

Traditional metrics such as BLEU (Papineni et al., 2002), which utilizes n-grams, and METEOR (Banerjee and Lavie, 2005), which considers an alignment using unigrams, rely on the textual similarity between the model's output and reference texts to produce evaluation scores. Additionally, BERTScore (Zhang et al., 2020), which measures semantic similarity, and generation-based metrics like Prism (Vernikos et al., 2022), rely on the textual similarity between the model' s output and reference texts to produce evaluation scores. Recent metrics, such as BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), measure the semantic similarity using language models, and they employ neural networks to enhance score generation. These metrics focus more on semantic understanding by aligning their assessments with human evaluations.

The Multidimensional Quality Metrics (MQM) framework (Burchardt, 2013; Mariana, 2014; Freitag et al., 2021a) is a detailed and flexible approach to further refine the task of evaluating translation quality. It assesses word-level errors, semantic accuracy, stylistic nuances, and cultural appropriateness. The MQM Core [3] includes 39 distinct error types, and the MQM Full [4] includes even more, offering a comprehensive translation quality assessment. In addition to error types, the MQM framework allows annotators to tag each error with

[3] https://themqm.org/the-mqm-typology/
[4] https://themqm.org/the-mqm-full-typology/

	Chat Domain	Human Evaluation Method	Evaluation Focus	Fine-grained Analysis	Language Pairs
WMT 2020 Chat Translation	Custom Service	Segment Rating + Document Context	Pronoun (*it*).	△	en⇔de
WMT 2022 Chat Translation	Custom Service	Adapted MQM*	Accuracy, Linguistic Conventions, Terminology, ... MT Hallucination, Source Issue.	△	en⇔de, en⇔fr, en⇔pt_br
CPCC	Custom Service, TV series	Customized	Preference, Coherence, Consistency, Fluency.	○	en⇔de, en⇔zh
CSA-NCT	Custom Service, TV series	Customized	Coherence, Speaker, Fluency.	○	en⇔de, en⇔zh
SML	Custom Service, TV series	Question-based	Coherence, Fluency.	○	en⇔de, en⇔zh
MQM-Chat Annotation	**Various** (news, sports, hobbies, daily life, social media, etc.)	MQM-Chat	Source Issue→**Disambiguation**, Consistency→**Dialogue Consistency**, Speaker→**Stylized Contents**, **Cultural Contents**, **Buzzwords and Loanwords**.	○	zh⇒en **ja**⇒en

Table 2: Comparison of previous studies across several dimensions: data domain, human evaluation method, evaluation focus, granularity of results, and language pairs studied.

a severity to indicate its impact on the translation. Recent developments in automatic MQM evaluation include xCOMET (Guerreiro et al., 2023), AutoMQM (Fernandes et al., 2023), and GEMBA-MQM (Kocmi and Federmann, 2023). AutoMQM and GEMBA-MQM leverage LLMs to automatically detect and label the errors with MQM error types and severity. Compared to existing metrics, we defined a benchmark that provides a multidimensional evaluation with error types specific to chat translation tasks.

2.2 Chat Translation Tasks

Chats frequently include slang, idiomatic expressions, and personalized styles, increasing the translation task's complexity (Baldwin et al., 2013; Eisenstein, 2013). While high accuracy is important when translating chats, preserving the nuances and special contents is sometimes even more crucial (Hovy, 2015; Salganik, 2020).

The first workshop that focused on chat translation was the Fifth Conference on Machine Translation (WMT20) (Barrault et al., 2020; Farajian et al., 2020), which laid the groundwork in this domain. This was followed by the Seventh Conference on Machine Translation (WMT22) (Kocmi et al., 2022; Farinha et al., 2022) and continued with the Ninth Conference on Machine Translation (WMT24) (Mohammed et al., 2024). The WMT shared tasks have primarily focused on customer service chats, which are relatively structured and standardized. The emphasis has been on evaluating the overall performance of chat translation models with a strong focus on syntax accuracy. The WMT2022 shared task began to address chat-specific issues, and Liang's team, as a continuation of WMT2020, presented improved chat translation models, highlighting the importance of coherence, fluency, and speaker personalities (Liang et al., 2021a,b, 2022).

Gradually, WMT and derivative studies have recognized the importance of source content issues and preserving the speaker's style in chat translations. Thus, MQM was adapted in the WMT2022 shared task; however, it was too broad with 31 error types, most of which focused on accuracy and relatively superficial analyses. Previous studies have utilized binary classification for chat translation with a specific focus on coherence (Li et al., 2022, 2023), which did not capture the complexity of chat translations effectively.

With these foundations, we have refined the evaluation process by differentiating the source issues in chat translations into ambiguity issues and cultural nuances issues such as buzzwords, and by emphasizing the importance of dialogue consistency. Additionally, we have de-emphasized grammatical accuracy, which is not always the highest priority in everyday conversations. To make MQM-Chat broadly applicable to general chats, we chose data that covers various topics, including news, sports, hobbies, daily life, and social media. Furthermore, we have included Japanese data, a language that has not been studied extensively in chat transla-

tion tasks. A comparison between our research and previous studies is presented in Table 2.

3 Multidimensional Quality Metrics for Chat Translation (MQM-Chat)

In this study, we define high-quality chat translation as maintaining accuracy while simultaneously capturing and conveying the speaker's personality, styles, and cultural nuances effectively. We refined the MQM framework and introduced customized categories that are specific to chat translation tasks.

3.1 Error Types

As mentioned previously, MQM-Chat focuses on seven error types: *Mistranslation*, *Omission or Addition*, *Terminology or Proper Noun Issues*, *Unnatural Style*, *Ambiguity and Disambiguation*, *Buzzwords or Loanwords Issues*, and *Dialogue Inconsistency*. The latter three error types are customized typologies for chat translation. The errors are evaluated with three levels of severity to provide a sufficiently detailed and accurate assessment. The mapping of error types between MQM-Chat and standard MQM is shown in Figure 1. Note that the relationship between the mapping blocks is not just an inclusion relationship because error types in MQM-Chat cover broader issues in chat translation tasks with specific descriptions and examples.

Mistranslation *Mistranslation* refers to fundamental inaccuracies in the translation process, including untranslated source segments, incorrect lexical choice or grammar that distorts the meaning, as well as undertranslation and overtranslation. These errors are critical because they directly impact the comprehensibility and accuracy of the translation.

Omission or Addition Missing source content (omission) or additional content not present in the source (addition) are considered to be *Omission or Addition* errors. Such errors can significantly mistake the intended message and disrupt the coherence of the translated text, which can result in misunderstandings.

Terminology and Proper Noun Issues *Terminology and Proper Noun Issues* are related to inaccuracies when translating specialized vocabulary, inherent terms, and proper nouns from the source text. Misinterpretations in this category can undermine the reliability of the translation, especially in professional and academic contexts. Note that this category does not include Internet terms, popular terms, newly created words, memes, and foreign words.

Unnatural Style *Unnatural Style* refers to translations that are grammatically correct but unnatural in the target language.

Ambiguity and Disambiguation The goal of this study is to retain the speaker-specific stylized contents and accurately translate them into their corresponding errors in the target language. *Ambiguity and Disambiguation* errors occur when the ambiguities or errors in the source text, such as typographical errors, omissions, unclear abbreviations, and erroneous punctuation, are not faithfully reflected in the translation. For example, the typos 知ってr and 队啊 shown in Table 1 are considered to be ambiguity errors. Deviations from this principle are considered errors, which highlights the need to translate the speaker-specific stylized content into corresponding errors in the target language. This error category emphasizes the importance of maintaining the authenticity of the source text, including its imperfections.

Buzzword or Loanword Issues *Buzzword or Loanword Issues* occur when such terms are not translated accurately according to their usage in both the source and target languages. This includes the incorrect translation of popular sayings, newly created words, Internet slang, and memes. For example, in Table 1, the Japanese meme 草 (representing laughter) and the Chinese Internet slang 鼠的 (meaning yes) are frequently mistranslated by existing translation models, capturing only the superficial aspects of their usage. This results in errors that obscure the source text's intended meaning and cultural nuance of the source text. Thus, if there is no corresponding term in the target language, the pronunciation should be retained and written in the target language.

Dialogue Inconsistency *Dialogue Inconsistency* occurs when translations fail to maintain consistency based on context, particularly when the speakers change in the chat. This can include inappropriate handling of demonstrative pronouns, personal references, or definite articles. For example, in Japanese and Chinese, the subject is frequently omitted when it has already appeared in the preceding context. As shown in Table 1, the subjects 你 (you) in the Chinese source and まどか (Madoka, she) in the Japanese source are omitted, which results in errors. Maintaining sufficient consistency

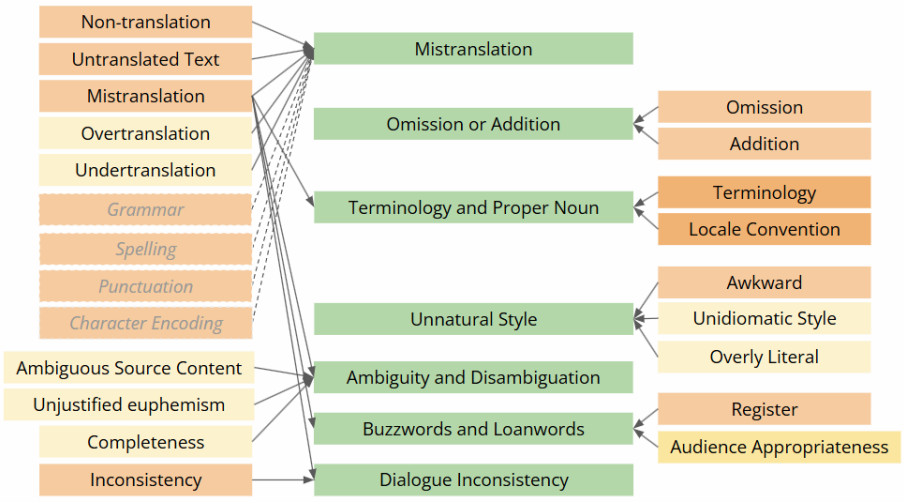

Figure 1: Mapping of error types in MQM-Chat (green) and MQM Core (orange) and used MQM Full (yellow). Blocks with deeper colors (*Terminology*, *Locale Convention*, and *Audience Appropriateness*) suggest that corresponding sub-categories are included and merged into MQM-Chat. Blocks with gray text (*Grammar*, *Spelling*, *Punctuation*, *Character Encoding*) are errors that are only marked if they totally interrupt the translation substantially. Note that the relationship between the mapping blocks is not simply an inclusion relationship because MQM-Chat error types cover broader issues in chat translation.

in dialogue is crucial to ensure coherence and avoid confusing the reader.

4 Experiments

Experiments were conducted to evaluate the effectiveness of MQM-Chat by translating chats from Japanese (ja) and Chinese (zh) into English (en) and assigning the MQM-Chat human annotations, MQM-Chat automatic annotations, and standard MQM automatic annotations.

4.1 Datasets

In the experiments, 200 short chats were selected from the Open 2ch Dialogue Corpus (Inaba, 2019) for the ja⇒en translations, which feature ambiguous content and popular sayings from Japan's online community 2channel. Sensitive content was excluded to avoid offensive data. Similarly, we selected 200 short chats from the LCCC-base dataset (Wang et al., 2020) for the zh⇒en translations. To provide an effective comparison and a broader range of chat contents, we included 100 long chats from BPersona-chat (Sugiyama et al., 2021; Li et al., 2022) for ja⇒en and 100 long chats from NaturalConv (Wang et al., 2021) for zh⇒en. Data statistics are listed in Table **??** in Appendix **??**.

4.2 Translation Models

Several translation models were considered in the experiments, including sentence-to-sentence transformers-based models (Vaswani et al., 2017), LLMs, and commercialized systems to generate the translation data. Specifically, they are GPT-4, LLaMA3 (70B-Instruct), DeepL, Facebook@WMT21 for zh⇒en and SKIM@WMT23 for ja⇒en. GPT-4 and LLaMA3 were used in zero-shot learning configurations (Romera-Paredes and Torr, 2015; Wang et al., 2019) with prompts designed on methodologies proposed by Hendy et al. (2023) and other recent studies (Farinhas et al., 2023; Peng et al., 2023). Detailed prompts and model parameters are listed in Appendix **??**.

4.3 Human Annotations

Six professional annotators proficient in Japanese and English and another six proficient in both Chinese and English were recruited through crowdsourcing. The annotators identified the translation errors and assigned the severity levels based on MQM-Chat specifications using Label Studio[5] (Tkachenko et al., 2020-2022). The annotators were provided with detailed guidelines to ensure sufficient consistency in the error labeling and severity assessment consistency. In addition, we manually reviewed the human annotation results and made necessary corrections to ensure the

[5] https://labelstud.io/

quality of the annotations. The reviewer examined the annotations, primarily focusing on whether the label and the severity of the annotations matched their definitions and whether the error span was overly broad. For example, if there was only a single error word but the error span contained the entire sentence, the reviewer would correct the span. Two reviewers familiar with MQM-Chat error types and proficient in Chinese, Japanese, and English checked the annotations afterward. Details about annotation tasks are presented in Appendix **??**.

To ensure the quality and safety of the dataset, the annotators were empowered to report any data containing extremely offensive or toxic content, which was implemented to avoid including highly toxic data in the experiments, thereby ensuring a more ethical and controlled evaluation process. The reported data were reviewed and excluded if deemed inappropriate to maintain the integrity and safety of the annotating task environment.

5 Results and Analysis

5.1 Error Distributions of MQM-Chat

As shown in Figure 2, we analyzed the error distributions of MQM-Chat annotations. The results demonstrate that, although the distribution of *Mistranslations* was skewed, MQM-Chat provided a varied distribution of errors across other categories. A possible reason for having more mistranslations could be that not all the translation models were specifically trained or fine-tuned on parallel zh-en or ja-en chat translation data.

When chat-specific errors occurred, MQM-Chat provided several insights. First, the ja⇒en translations generally exhibited more errors than the zh⇒en translations. In terms of the chat length, *Mistranslations* and *Unnatural Style* errors tended to occur frequently in the long chat translations compared to the short chats in both language pairs. Furthermore, *Omissions and Additions* were common in long chats, especially in the ja⇒en translations. In contrast, *Ambiguity* and *Buzzword Issues* appeared in short chats more frequently. In addition, *Dialogue Inconsistency* issues were found to be persistent across both the short and long chats, regardless of the source language.

For the ja⇒en translations, regardless of the chat length, the most frequent errors included *Omissions or Additions*, *Unnatural Style*, and *Dialogue Inconsistency* issues. We found that *Ambiguity and Disambiguation* errors were less common in the ja⇒en translations; however, they occurred at a similar frequency to other errors in the zh⇒en translations. This could be because Chinese people frequently omit punctuation when chatting, thereby leading to a more even distribution of ambiguous content in the Chinese segments. In addition, *Terminology and Proper Noun Issues*, *Ambiguity and Disambiguation*, and *Buzzword or Loanword* errors occurred more frequently in the short Japanese chats than in the long chats. Overall, the findings demonstrate that MQM-Chat provides valuable insights into the overall trends across languages and offers a deeper understanding of the translation challenges associated with different source languages.

In terms of the translation models, GPT-4 generated considerably fewer *Mistranslations* in the short chat translation tasks than the other three models. However, the amounts were similar for the long chat translation tasks. GPT-4 produced more *Buzzword or Loanword Issues* in long chats, and Llama3 struggled with dialogue consistency in the short chats and produced more *Terminology or Proper Noun Issues* and *Unnatural Style* problems in the short Japanese chat translation tasks. Furthermore, DeepL produced more *Terminology and Proper Noun Issues* in the zh⇒en translations. The results of MQM-Chat help us better understand the strengths and weaknesses of different models, thereby offering pathways to improve such models. For example, leveraging the glossary function in DeepL could help reduce terminology errors.

5.2 Errors Relabeled by MQM-Chat

To demonstrate the superiority of MQM-Chat over existing evaluation metrics in terms of capturing chat-specific translation errors, we compared the behavior of MQM-Chat and standard MQM on the same datasets. Here, we assigned the standard MQM human annotations, including *Accuracy* (*Addition*, *Mistranslation*, *Omission*, *Untranslated Text*), *Fluency* (*Character Encoding*, *Grammar*, *Inconsistency*, *Punctuation*, *Register*, *Spelling*), *Locale convention* (*Currency*, *Date*, *Name*, *Telephone*, or *Time Format*), *Style* (*Awkward*), *Terminology* (*Inappropriate for Context*, *Inconsistent Use*), *Non-translation* and *Others* (Freitag et al., 2021b). The standard MQM human annotations were applied to 25% of the data as samples.

To determine the extent of the actual effect of shifting to MQM-Chat, we evaluated the percentage of standard MQM's errors that were labeled as

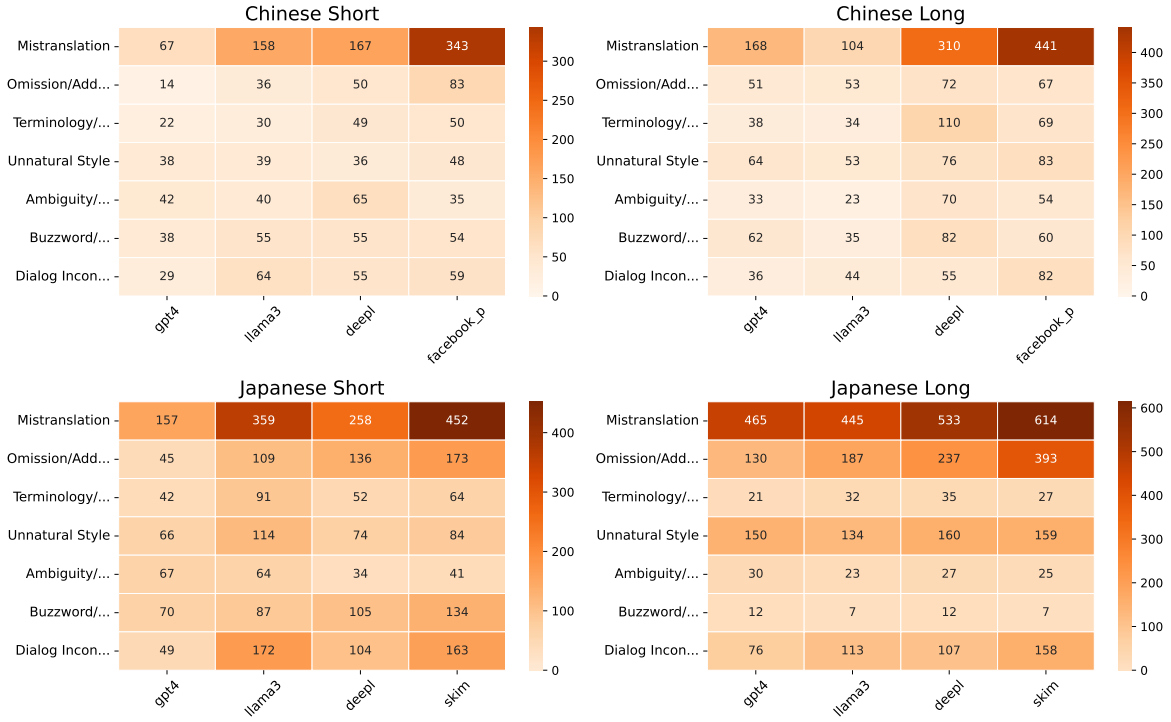

Figure 2: Heatmaps of the error numbers in MQM-Chat human annotations. Darker colors indicate higher numbers.

Data	Model	*zh⇒en* Relabeled (%)	*zh⇒en* Chat-spec (%)	*ja⇒en* Relabeled (%)	*ja⇒en* Chat-spec (%)
Short	**GPT-4**	39 (50.65%)	23 (29.87%)	92 (57.86%)	41 (25.79%)
	LLaMA3	46 (48.94%)	29 (30.85%)	201 (74.17%)	80 (29.52%)
	DeepL	46 (52.87%)	18 (20.69%)	172 (81.13%)	77 (36.32%)
	NMT	94 (67.63%)	35 (25.18%)	329 (86.58%)	120 (31.58%)
Long	**GPT-4**	16 (27.59%)	3 (5.17%)	33 (25.19%)	10 (7.63%)
	LLaMA3	21 (26.25%)	10 (12.50%)	69 (30.53%)	17 (7.52%)
	DeepL	71 (49.65%)	34 (23.78%)	92 (43.81%)	26 (12.38%)
	NMT	128 (55.41%)	48 (20.78%)	278 (61.50%)	49 (10.84%)

Table 3: Percentage of errors labeled with standard MQM human annotations were relabeled with MQM-Chat human annotations. **Relabeled (%)** represents the number and percentage of errors being re-labeled as other error types in MQM-Chat human annotations. **Chat-spec (%)** indicates the percentage of errors relabeled to chat-specific error types, such as *Ambiguity*, *Buzzword Issues*, and *Dialogue Inconsistency*.

other types by MQM-Chat. The results are shown in Table 3. The results suggest that MQM-Chat can successfully recognized at least 25.19% and at most 86.58% of the errors. Half of those errors were labeled as chat-specific errors with the help of MQM-Chat. Especially for the short chat, the percentage of errors labeled as chat-specific types was higher than that for the long chat, which also meets the nature that short data from LCCC-base and Open 2ch Dialogue Corpus include more source-side ambiguity, buzzword, and inconsistency issues.

Three examples are listed according to the three chat-specific error types in Table 4. In the *Ambiguity and Disambiguation* example, misspelling 那地 as 那滴 caused the translation model to generate an error, which was labeled as *Ambiguity and Disambiguation* according to MQM-Chat, while standard MQM marked this error as a *Mistranslation*. The second example shows an error caused by the Internet slang 人艰勿拆, which was labeled as a *Buzzword Issue* by MQM-Chat but a *Mistranslation* by standard MQM. In the third example, MQM-Chat labeled a sentence with referential problems as *Dialogue Inconsistency* rather than *Mistranslation*.

From the comparison with standard MQM, we believe that MQM-Chat differs from existing evalu-

Example 1	
source	我在那滴 (a typographical error of “那地”)吃的饭。。
NMT output	I had my meals in that drop....
(possible reference)	theere (a typographical error of “there”)
Standard MQM	*Mistranslation* - Critical
MQM-Chat	***Ambiguity and Disambiguation* - Major**
Example 2	
source	...我只是为了凹造型人艰勿拆
DeepL output	I just for the sake of the shape of the people hard not to break down
(possible reference)	Ren Jian Wu Chai (Chinese transliteration) or life is already so hard or arduous, so don’t judge me. (the meaning)
Standard MQM	*Mistranslation* - Major
MQM-Chat	***Buzzword or Loanword Issues* - Major**
Example 3	
source	...結婚して早く家を出ろって母がうるさくて。 - そうだったのかぁ。うち(*1)とは逆だね。うちは一緒にいてほしいみたいだよ。(*2)
NMT output	...my mother insisted that I get married and leave the house as soon as possible. - I see, it’s the opposite of my house(*1). I want you to stay with me.(*2)
(possible reference)	(*1) my family, (*2) She want me to stay with her.
Standard MQM	(*1) *Mistranslation* - Major, (*2) *Mistranslation* - Critical
MQM-Chat	(*1) *Mistranslation* - Major, (*2) ***Dialogue Inconsistency* - Major**

Table 4: Examples of errors being labeled as *Mistranslations* by standard MQM annotations that were classified into chat-specific labels such as *Ambiguity and Disambiguation*, *Buzzword or Loanword Issues*, and *Dialogue Inconsistency* in MQM-Chat annotations. **Standard MQM** and **MQM-Chat** columns show the judgements on the MT output by each annotation criteria ([label] - [severity]).

	zh⇒en			*ja⇒en*		
	Span			**Span**		
Data	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Short	52.24	64.99	54.03	55.93	42.26	43.65
Long	33.97	43.56	33.45	38.74	18.63	22.30
	Span+Label			**Span+Label**		
Data	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Short	24.26	27.47	24.38	20.28	15.81	16.28
Long	14.11	16.54	13.55	10.75	5.45	6.49
	Span+Severity			**Span+Severity**		
Data	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Short	31.35	37.54	32.11	29.67	21.97	23.05
Long	16.69	20.64	16.33	17.69	8.82	10.39

Table 5: Average precision, recall and F1 scores of MQM-Chat automatic annotations having span overlap, span and label overlap, span and severity overlap with human annotations.

ation benchmarks because it can locate and analyze problems specific to chat translation precisely.

6 Automatic MQM-Chat Annotations

We implemented MQM-Chat automatic evaluation based on the GEMBA-MQM (Kocmi and Federmann, 2023) prompt by replacing the description of the error types of standard MQM with the error types of MQM-Chat. Three chat translations with MQM-Chat annotations were provided as examples for few-shot learning. The prompt is shown in Appendix **??**. Using human annotations as the golden standard, we calculated the pairwise accuracy and Pearson agreements for the automatic evaluation. The 79.17% pairwise accuracy and 0.774 Pearson correlation values demonstrate that the auto annotations of MQM-Chat agree with human rankings.

We also reviewed the error spans of the proposed MQM-Chat to investigate whether its annotations can also analyze errors in detail. The distribution of the error numbers is shown in Figure 3. Unlike human annotations, automatic annotations focus on *Mistranslation* and *Unnatural Style* in all four cases. The number of *Unnatural Style* errors was even greater than that of *Mistranslations* for the LLMs’ zh⇒en translations. Additionally, fewer errors were observed in ja⇒en than in zh⇒en, with always more *Unnatural Style* errors, which is the opposite of the human annotations, The findings may be related to the limited amount of Japanese chat data, compared with the Chinese chat data, in GPT’s training dataset.

Span-level accuracy is based on the methodology employed in AutoMQM; however, a more flexible approach is adopted by considering spans as overlapping if they mostly align rather than strictly

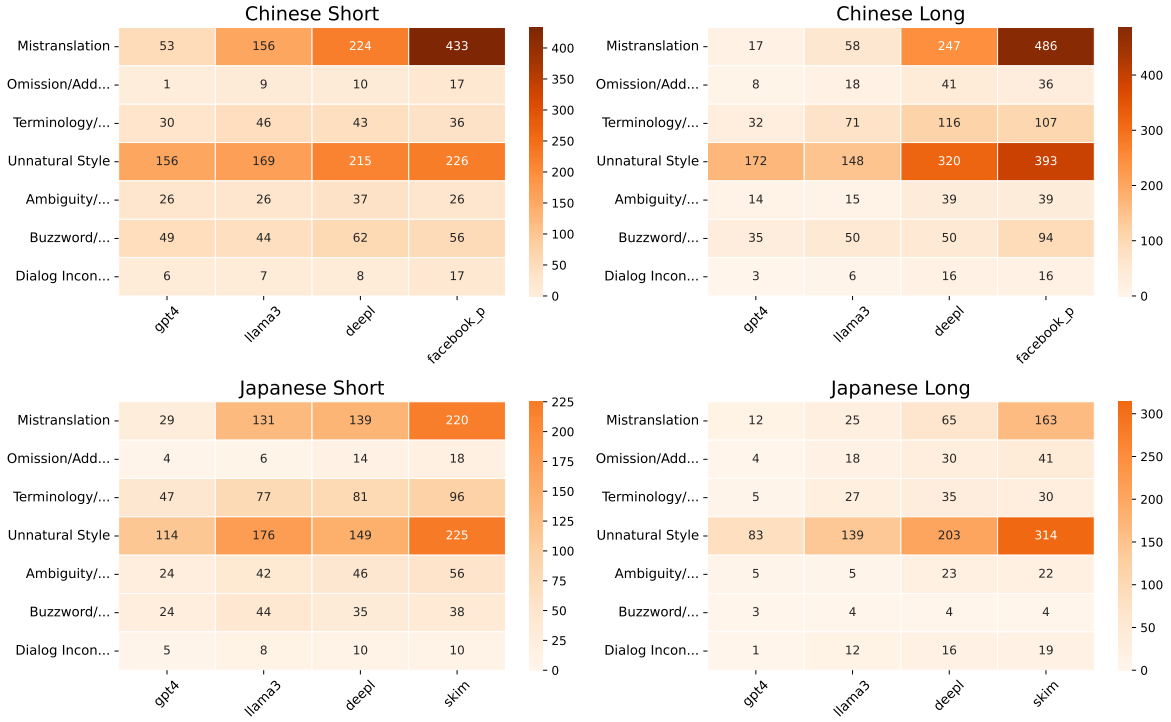

Figure 3: Heatmaps of error numbers in MQM-Chat auto annotations. Darker colors indicate higher numbers.

counting overlapping characters. Here, we calculated precision, recall, and F1 scores for three scenarios: span overlap, both span and label overlap, and both span and severity overlap, as shown in Table 5. The results demonstrate that MQM-Chat auto annotations have limited overlap with human annotations, particularly with the long chat translation tasks. We consider that the current few-shot prompting approach with GPT-4 is insufficient to fully support the automatic evaluation of MQM-Chat. However, this is understandable, given that MQM-Chat is a new evaluation metric. We plan to optimize MQM-Chat automatic annotations by selecting open-source LLMs and utilizing human annotations for fine-tuning in the future.

7 Conclusion

To address the lack of evaluation metrics for chat translation tasks, this study has proposed MQM-Chat and evaluated its effectiveness through a series of experiments. By analyzing the error distribution, we find that MQM-Chat is suitable for qualifying chat translations and can successfully identify the weakness of the experimented translations. Looking at the errors that were relabeled as other error types by MQM-Chat, we consider that MQM-Chat can provide a more nuanced classification of errors, especially for chat-specific issues.

In addition, we explored MQM-Chat automatic annotations by few-shots learning on GPT-4 with GEMBA-MQM's prompt. It agrees with the system rankings annotated by human annotators but cannot fully overlap with the spans in human annotations. However, although it currently needs further refinements, we still consider that it can serve as a basic reference.

In this study, to address the complexities of translating everyday chat conversations, we selected data that were rich in slang and ambiguous contents, which inherently increases the difficulty of the translation tasks. In the future, we plan to evaluate MQM-Chat on chat translation data from other domains, such as custom service, to determine if the results have the same characteristics. In addition, we plan to enhance the implementation of automatic MQM-Chat, thereby enabling it to serve as an effective evaluation reference for chat translation tasks. Ultimately, we hope that the MQM-Chat can be used as a valuable evaluation benchmark for chat translation tasks to facilitate good performance in translation tasks involving chats and other informal content translations.

Limitations

With data limited to translations from Chinese and Japanese to English, the experimental results obtained in the current study are relatively narrow. Thus, in the future, we plan to extend MQM-Chat to more language pairs and bidirectional translations to better understand chat translation across various languages. In summary, we consider that MQM-Chat has laid a solid foundation for this type of research, opening up many potential directions to improve and expand chat translation evaluations.

Ethical Considerations

The crowdsourcing experiments conducted in this study adhered to stringent ethical guidelines to ensure participant privacy and data protection. The experiments deliberately avoided collecting any personally identifiable information from the participants. No restrictions or enforcement of specific work hours were imposed upon the participants, thereby eliminating undue influence or coercion. Given the absence of personal data collection and voluntary participation, the data were not subject to an ethics review at the organization. Consequently, the data collection procedures used in this study adhered to the ethical standards and regulations governing acceptable research practices.

Acknowledgements

This work was supported by JST (the establishment of university fellowships towards the creation of science technology innovation) Grant Number JPMJFS2102, JST SPRING Grant Number JPMJSP2114, JSPS KAKENHI 22H00524, JST CREST Grant Number JPMJCR20D2 and JST Moonshot R&D Grant Number JPMJMS2011-35 (fundamental research). The crowdsourcing was supported by Crowdworks (https://crowdworks.jp/). The annotation was supported by Label Studio (https://labelstud.io/), especially thanks for providing the academic version.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. 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.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, He-

lena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021b. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762.

Michimasa Inaba. 2019. A example based dialogue system using the open 2channel dialogue corpus. In *Proceedings of SIG-SLUD-B902-33*, pages 129–132.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 general translation task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 128–136, Singapore. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2023. An investigation of warning erroneous chat translations in cross-lingual communication. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 10–16, Nusa Dua, Bali. Association for Computational Linguistics.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard, and Kentaro Inui. 2022. Chat translation error detection for assisting cross-lingual communications. In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 88–95, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021a. Modeling bilingual conversational characteristics for neural chat 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 5711–5724, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. Scheduled multi-task learning for neural chat translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4375–4388, Dublin, Ireland. Association for Computational Linguistics.

Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2021b. Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 67–79, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. Association for Computational Linguistics.

Gennady Medvedev. 2016. Google translate in teaching english. *Journal of teaching English for specific and academic purposes*, 4(1):181–193.

Wafaa Mohammed, Sweta Agrawal, Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C Farinha, and José G. C. De Souza. 2024. Findings of the WMT 2024 shared task on chat translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 701–714, Miami, Florida, USA. Association for Computational Linguistics.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. 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.

Sumant Patil and Patrick Davies. 2014. Use of google translate in medical communication: evaluation of accuracy. *Bmj*, 349.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR.

Matthew J Salganik. 2020. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. *Preprint*, arXiv:2109.05217.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2024. Llama: Open and efficient foundation language models. `https://ai.facebook.com/blog/large-language-models-llama-3`.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

David C Uthus and David W Aha. 2013. Multiparticipant chat analysis: A survey. *Artificial Intelligence*, 199:106–121.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level mt metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. *Preprint*, arXiv:2103.02548.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR 2020*.