An Investigation of Warning Erroneous Chat Translations in Cross-lingual Communication

Yunmeng Li[1] **Jun Suzuki**[1,3] **Makoto Morishita**[2] **Kaori Abe**[1*] **Kentaro Inui**[4,1,3]

[1]Tohoku University [2]NTT [3]RIKEN [4]MBZUAI

li.yunmeng.r1@dc.tohoku.ac.jp

Abstract

Machine translation models are still inappropriate for translating chats, despite the popularity of translation software and plug-in applications. The complexity of dialogues poses significant challenges and can hinder cross-lingual communication. Instead of pursuing a flawless translation system, a more practical approach would be to issue warning messages about potential mistranslations to reduce confusion. However, it is still unclear how individuals perceive these warning messages and whether they benefit the crowd. This paper tackles to investigate this question and demonstrates the warning messages' contribution to making chat translation systems effective.

1 Introduction

Globalization has led to the popularity of neural machine translation (Bahdanau et al., 2014; Vaswani et al., 2017; Gehring et al., 2017). Applications like Google Translate[1] and DeepL[2] have become essential tools in people's lives (Medvedev, 2016; Patil and Davies, 2014). Chat software such as WeChat and LINE also integrates built-in translation features to facilitate cross-lingual communication. Plug-in translating applications like UD Talk[3] and Hi Translate[4] have become popular as well with the rise of online communication.

However, while machine translation technologies have demonstrated sound performance in translating documents (Barrault et al., 2019, 2020; Nakazawa et al., 2019; Ma et al., 2020; Maruf and Haffari, 2018), current methods are not always suitable for translating conversations (Uthus and Aha, 2013), especially colloquial dialogues such as chats (Läubli et al., 2018; Toral et al., 2018;

Farajian et al., 2020; Liang et al., 2021a). When a translation system generates erroneous translations, people unable to read the other language may not recognize such errors, leading to confusion.

Achieving a perfect error-free chat translation system is challenging due to the unique characteristics of chat (Tiedemann and Scherrer, 2017; Maruf et al., 2018; Liang et al., 2021a,b), making it impractical to aim for perfection. Instead, a viable alternative approach is to enhance translation software by providing warnings about possible mistranslations to reduce confusion. However, the perception and effects of such warning messages remain unclear. To investigate this, we proposed to provide a warning message for erroneous translations during the cross-lingual chat and conducted a survey to explore how such warnings help people communicate. The survey design is shown in Figure 1. Participants engage in a simulated cross-lingual chat scenario, where they have to select the most reasonable response from three options. Whenever a translation error occurs, a warning message is displayed. At the end of the chat, participants answer corresponding questions regarding their perceptions of the warning messages.

We conducted the survey and collected responses through crowdsourcing. The results indicate that warning messages (1) are helpful in cross-lingual chats and (2) potentially encourage users to change their chat behavior. Moreover, the survey reveals the crowd's desired features for the warning messages. This is the first study of its kind to explore the impacts of warning users about erroneous translations in cross-lingual chat. The findings are valuable for developing an assistant function that detects and warns users of erroneous chat translations.

2 Related Work

Previous studies have pointed out the potential benefits of incorporating machine translation in chat, despite its imperfections (Uthus and Aha, 2013).
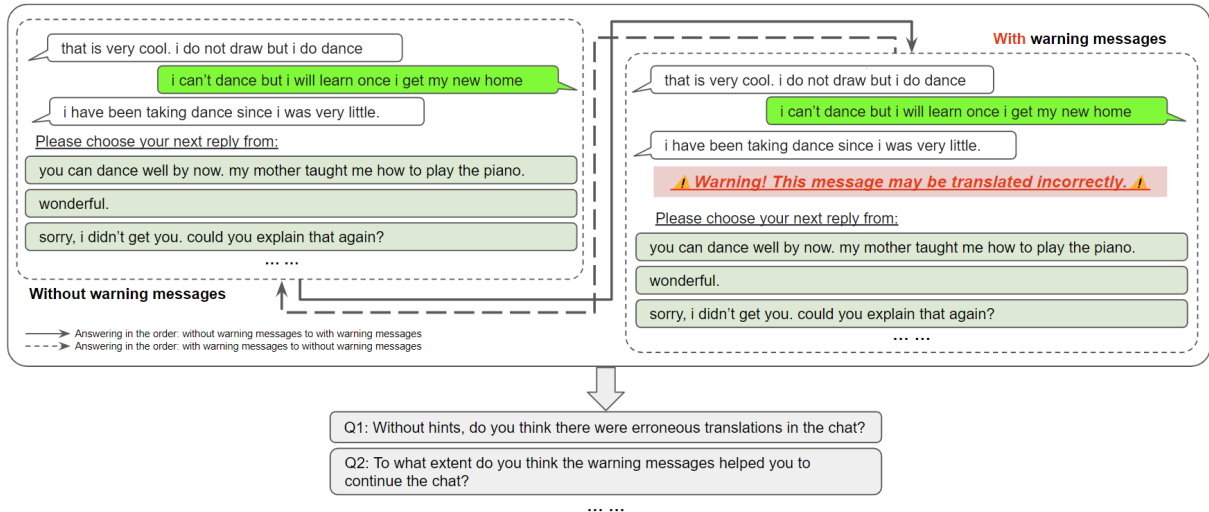

Figure 1: An illustration of the designed survey. Participants will engage in two rounds of chat in the survey: one without warning messages (left) and one with warning messages (right). The content and response options are the same in both rounds. The order of the two rounds, either "without-with" (solid line) or "with-without" (dotted line), will be randomly assigned to participants.

Several researchers have trained models using different methods to enhance chat translation performance (Maruf et al., 2018; Farajian et al., 2020; Liang et al., 2021a). However, features such as ambiguity, omissions, and multi-speakers make it challenging to improve translation accuracy in chat (Tiedemann and Scherrer, 2017; Liang et al., 2021a,b). In contrast to existing studies of training chat translation models, we focus on acknowledging the imperfect nature of machine translation (Uthus and Aha, 2013) and aim to enhance people's experience of chat translation through an alternative approach. We propose the warning message of erroneous translation and thus improve people's experience in cross-lingual chat. A chat translation error detector discussed in a recent study provides a binary assessment of the coherence and correctness of chat translations (Li et al., 2022b). If the error detector's predictions are transformed into warning messages, our survey could be instrumental in assessing the error detector's practical effectiveness. To the best of our knowledge, the study is the first to investigate the crowd's acceptance of such chat translation error detection tasks.

3 Survey Design

We propose an alternative strategy to improve translation software's performance by integrating cautionary alerts for potential mistranslations to reduce confusion. We designed a warning message and executed a survey to evaluate its effectiveness. Fig-

ure 1 illustrates the survey process, including two simulated chat rounds: one devoid of warning messages and the other incorporating them.

3.1 Simulated Cross-lingual Chat Scenarios

Since dynamic real-time chats are relatively uncontrollable and high-cost, we simulated a chat scenario with a foreign partner based on chat data from Persona-chat (Zhang et al., 2018). In the simulation, participants are presented with three initial chat turns as historical chat logs at the beginning.Participants choose the most contextually fitting response from the three provided options each time their scripted partners respond iteratively. To explore the cognitive processes of individuals lacking proficiency in a foreign language, we operated under the assumption that participants would receive translated messages generated by the machine translation system from their partners. Hence, all texts within the survey are presented to participants in their native language.

3.2 Chat Data

We prepared the simulated scenarios with the Persona-chat dataset, containing multi-turn chat data about various personality traits with assumed personas in English. To ensure the quality of the data, we eliminated incoherent and unnatural chat data from Persona-chat through crowdsourcing at Amazon Mechanical Turk [5]. We defined "inco-

---

[5] https://requester.mturk.com/

11

herence" as questions being ignored, the presence of unnatural topic changes, one speaker not addressing what the other speaker said, responses appearing to be out of order or generally difficult to follow. We scored each chat according to the workers' answers and selected 6 of $1,500$ chats marked as accurate and coherent by at least seven of the ten workers. The chosen chats were used as the base of the simulated scenarios in the survey.

Similarly, we required proficient English speakers to continue the chat with given personas and topics from Persona-chat for other branching options and extended chats triggered by the options.

3.3 Erroneous Translations

To provide the chat data that were supposed to be erroneous translations, we translated the prepared chat data with a low-quality machine translation model that achieved a considerably low BLEU score (Papineni et al., 2002) of 4.9 on the English-Japanese chat translation evaluation dataset BPersona-chat (Li et al., 2022a). Consequently, we transformed the low-quality translations twenty times through Google Translate into different languages and finally translated them back to the source language of the survey. To ensure the final translations could serve as erroneous translations, we manually confirmed that the texts included significant syntax issues, incorrect emotional expressions, incoherence, or other errors that led to confusion. We designed that at least one of the three turns of the simulated chat would include erroneous translations. We required proficient English speakers to continue the chat based on the erroneous translations to prepare the extended chat.

3.4 Warning Messages

We designed the warning message to notify participants of erroneous translations in the chat. When the current text is assumed to be the erroneous translation, participants are presented with a warning message alerting them of the mistranslation, as shown in Figure 1. We structured the warning messages into two types since receiving and sending are both essential in a conversation. One type alerts participants of erroneous translations in the messages they received, while the other type indicates potential errors in the last message they sent.

3.5 Corresponding Questions

After the chat, participants are asked to answer if they notice erroneous translations without hints. If participants answer yes, they rate their experience on two Likert Scale questions (Joshi et al., 2015; Nemoto and Beglar, 2014). The first question assesses the extent to which the errors prevented them from continuing the chat, while the second question asks to what extent they could grasp exactly where the erroneous translations were in the message. Participants will use 1-5 to score their perceptions, with higher numbers indicating a greater awareness or understanding of the errors.

Participants must also rate on a Likert Scale question the extent to which they think the warning helped them continue the chat. Further, they check the plural options of additional features they find helpful if added to the warnings. Selectable features include: *indicating the correctness rate of the translation, providing alternative translation suggestions, showing specific errors in the translation,* and *suggesting the emotion of their partner.*[6]

4 Crowdsourcing Experiments

We prepared the survey in English, Chinese, and Japanese to observe the possible difference between languages. Professional translators translated the data from English to Chinese and Japanese to ensure quality. We prepared three sets of chat data for each type of warning message and two types of warnings; hence, we provided six sets of chat and collected the responses through crowdsourcing. We provided instructions for participants on how the chat would be presented and what they should do to attend the chat at the beginning of the task. Participants would be acknowledged that (1) their partner would speak to them in a language other than their native language, (2) the system would translate their partners' messages and the chat would only be presented in their language, (3) they would read the chat log and choose the most reasonable of the three options, (4) the message sent to them would be displayed on the odd-numbered lines, and their answer would be displayed on the even-numbered lines.

To minimize any possible influence of showing warnings first or later, we provided each chat in two orders. Participants answer either without warning messages first or with warning messages first. At the round of warning messages, we would explain the role of warning messages to participants and inform them that they could refer to the warnings

---

[6]Participants can fill in their comments or skip if they do not have any specific wanting features.

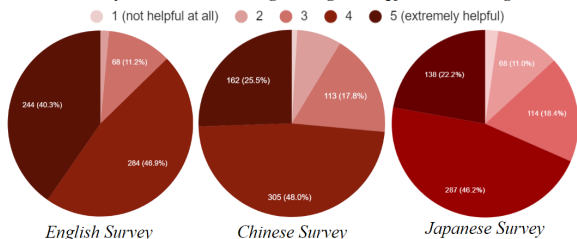How much do you think the warning message is helpful in cross-lingual chats?

1 (not helpful at all) 2 3 4 5 (extremely helpful)

English Survey   Chinese Survey   Japanese Survey

Figure 2: The responses to how participants think the warning messages helped them continue the chat.

changed choices before the turn   changed choices at the same turn
did not change choices at the same turn

Figure 3: The results that whether participants changed their choices with the help of warning messages.

to help them make choices.

We invited at least 50 participants for each order and ensured they could not join both orders through the crowdsourcing platforms' features. Crowd-workers were unaware of the fact that there were two orders, and they did not know which order they would join. Ultimately, we invited at least 100 participants for each set of chats.

The surveys were conducted on Amazon Mechanical Turk[7] for English participants, WenJuanXing[8] for Chinese participants, and CrowdWorks[9] for Japanese participants. Workers participated anonymously and were informed that the results would be used for academic purposes. Classification rounds were held in advance for efficiency.

5 Results and Analysis

Under the different policies of crowdsourcing platforms, we finally gathered 604 English, 635 Chinese, and 621 Japanese responses. Figure 2 displays the overall summaries. Around 70% of participants across three languages rated the warning messages as *"4 - helpful"* or above in the chat. Most participants view the warning messages as helpful in cross-lingual chats, aligned with Likert Scale analysis (Amidei et al., 2019).

With or without warning messages   The results of *"Without hints, do you think there were erroneous translations in the chat"* based on the order in which participants answered the survey are listed in Table 1. The percentages of noticing erroneous translations without hints remain consistent, regardless of participants answering with warning messages first or after. Hence, we conclude that the impact of answering orders on the crowds appears minimal. Moreover, considering a score greater or equal to 4 suggests the positivity of a Likert
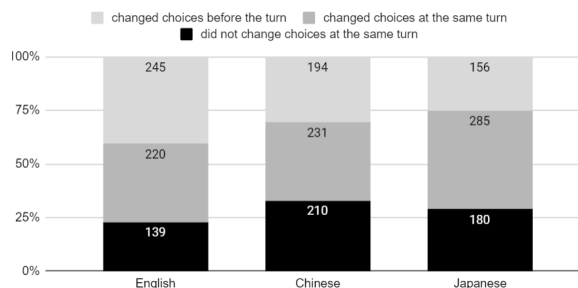
Scale question, we conclude that most participants who noticed erroneous translations also considered those errors as obstacles.

It is worth noting that while the English and Chinese results are relatively similar, Japanese results differ slightly. The recognition of erroneous translations without hints is notably lower in Japanese than in English and Chinese contexts. Participants' feedback suggests this may be related to Japanese linguistic specificity in "omission." Participants considered erroneous translations as omissions, aligning with Japanese conversational patterns where subjects or objects are often omitted. The warning messages helped them realize that the expression was not omitted but errors for the better continuation of the chat.

Additionally, English and Chinese participants also remarked that the warnings clarified unusual expressions as translation errors rather than humor or slang. The feedback helped state the usefulness of warning messages and the consideration for future differentiation between translation errors and humorous terms or buzzwords.

Impact of warning messages on modifying user's chat behavior   We analyzed participants' choices in relation to warning messages, categorizing them into three cases: (1) entered the same scenario in both the round with warnings and the round without warnings and did not change their choices, (2) entered the same scenario in both rounds and changed their choices, and (3) did not change their choices due to entering other branches in advance. We believe that the first case demonstrates that participants were not influenced by warnings, while the second case shows that they were influenced. In the third case, although it is impossible to compare whether participants changed their choices in the same scenario since they changed earlier, we still view it as an indirect influence due to the equiv-

	Without Warning Messages First					**With Warning Messages First**				
English	**Noticing mistranslations without hints**					**Noticing mistranslations without hints**				
	234 of 303 (77.2%)					234 of 302 (77.4%)				
	Considering mistranslations to be barriers					**Considering mistranslations to be barriers**				
	Score=1	*Score=2*	*Score=3*	*Score=4*	*Score=5*	*Score=1*	*Score=2*	*Score=3*	*Score=4*	*Score=5*
	2	11	56	126	39	5	17	55	108	49
Chinese	**Noticing the erroneous translations without hints**					**Noticing the erroneous translations without hints**				
	228 of 325 (70.2%)					241 of 310 (77.7%)				
	Considering mistranslations to be barriers					**Considering mistranslations to be barriers**				
	Score=1	*Score=2*	*Score=3*	*Score=4*	*Score=5*	*Score=1*	*Score=2*	*Score=3*	*Score=4*	*Score=5*
	2	26	45	112	53	2	26	62	115	36
Japanese	**Noticing the erroneous translations without hints**					**Noticing the erroneous translations without hints**				
	175 of 321 (54.5%)					158 of 300 (52.7%)				
	Considering mistranslations to be barriers					**Considering mistranslations to be barriers**				
	Score=1	*Score=2*	*Score=3*	*Score=4*	*Score=5*	*Score=1*	*Score=2*	*Score=3*	*Score=4*	*Score=5*
	3	21	29	89	33	1	17	29	86	25

Table 1: The results of the questions about noticing erroneous translations without hints in the two different answering orders. Participants who answered yes to the question continued to rate the extent they considered the erroneous translations to be barriers in the chat. The higher the score was, the more confused the participant felt.
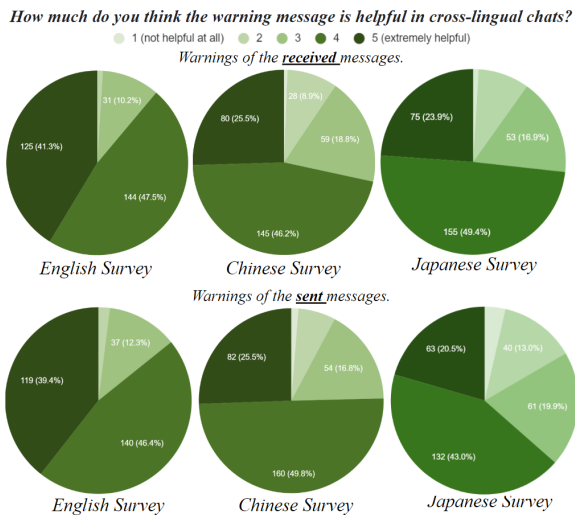

Figure 4: The responses to how participants think the warnings of the **received/sent** messages helped them continue the chat.

Figure 5: The results about expected additional features to the warning messages.

alence between having no warning messages and having no erroneous translations. Indeed, 103 participants stated they changed their choices as they ensured there were no erroneous translations.

Survey results shown in Figure 3 indicate that approximately 25% participants remained unchanged, while about 75% changed their choices, either directly or indirectly, due to the warning messages. We confirm that the participants were genuinely influenced by warning messages and participated in the subsequent feedback.

Warnings on the received messages or the sent messages The collected responses of different types of warning messages are summarized in Fig-
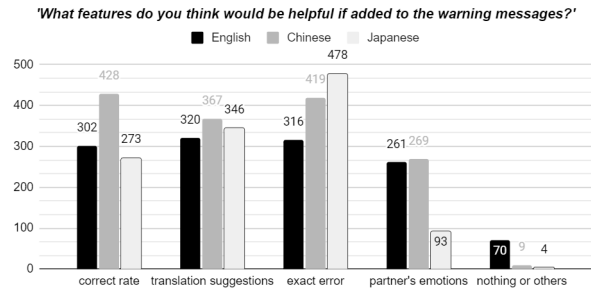
ure 4. Regardless of whether the warning messages indicated translation errors in the message received or sent, over 60% of the participants found the warning messages helpful (rating with a score-4 or higher) in all three languages.

Expected features of the warning message The results of expected additional information of the warning message are presented in Figure 5.

Chinese and Japanese participants showed a greater expectation for warning messages to indicate the exact error of their partners' messages. In addition, Chinese participants prefer to know the correct rate. Feedback from participants indicated that the correctness rate would better assist them in determining whether they needed to reinterpret. Japanese participants consider having other translation suggestions as references. English survey participants voted on all the listed features on average, but knowing their partner's emotions were still lower than others. In summary, to enhance the warning messages, the focus may better be on

highlighting the exact errors in the translations.

6 Conclusions

We conducted a survey to investigate the effectiveness of warning about possible mistranslations in chat as an alternative approach to enhance the experience of cross-lingual communication. Through crowdsourcing, we collected responses and concluded that such warning messages are helpful. By comparing the participants' choices with and without warning messages, we found that the warning messages did encourage participants to change their behaviors. We also found the crowd expects the warning message to (1) show the specific error in the translation, (2) indicate the correctness rate of the translation, and (3) provide alternative translation suggestions.

This survey is the first to explore the effects of warning about erroneous translations in cross-lingual chat, providing valuable insights for developing an assistant function that detects and warns people of erroneous chat translations.

Limitations

During the survey design phase, diligent measures were taken to minimize potential leading effects on the participants' judgment by randomly switching the order and neutralizing the questioning style. Despite the conscientious efforts, we must acknowledge the inherent challenges in completely eliminating all influences on the people who participated in the survey. With this realization, we recognize the need for further optimization to guarantee the fairness and validity of the responses. Refinement is warranted to minimize the biases further.

Ethics

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

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

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

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.

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.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional se-

quence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

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, Ryoko Tokuhisa, Brassard Ana, and Inui Kentaro. 2022a. Bpersona-chat: A coherence-filtered english-japanese dialogue corpus. In *Proceedings of NLP2022*, pages E7–3.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard, and Kentaro Inui. 2022b. 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, 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.

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.

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.

Tomoko Nemoto and David Beglar. 2014. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, pages 1–8.

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.

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.

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.

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.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.