Three Directions for the Design of Human-Centered Machine Translation

Samantha Robertson
University of California, Berkeley
samantha_robertson@berkeley.edu

Wesley Hanwen Deng
University of California, Berkeley
wesley1016@berkeley.edu

Timnit Gebru
Black in AI
timnit@blackinai.org

Margaret Mitchell
margarmitchell@gmail.com

Daniel J. Liebling
Google
dliebling@google.com

Michal Lahav
Google
mlahav@google.com

Katherine Heller
Google
kheller@google.com

Mark Díaz
Google
markdiaz@google.com

Samy Bengio
Google
bengio@google.com

Niloufar Salehi
University of California, Berkeley
nsalehi@berkeley.edu

Abstract

As people all over the world adopt machine translation (MT) to communicate across languages, there is increased need for affordances that aid users in understanding when to rely on automated translations. Identifying the information and interactions that will most help users meet their translation needs is an open area of research at the intersection of Human-Computer Interaction (HCI) and Natural Language Processing (NLP). This paper advances work in this area by drawing on a survey of users' strategies in assessing translations. We identify three directions for the design of translation systems that support more reliable and effective use of machine translation: helping users craft good inputs, helping users understand translations, and expanding interactivity and adaptivity. We describe how these can be introduced in current MT systems and highlight open questions for HCI and NLP research.

1 Introduction

Users often do not have enough information about the capabilities and limitations of machine learning models in order to use them effectively, safely, and reliably (Green and Chen, 2019; Pasquale, 2015). One approach to this problem is to provide additional information to users, such as explanations of model behavior (Liao et al., 2020; Lai and Tan, 2019; Rader et al., 2018). However, it is not clear

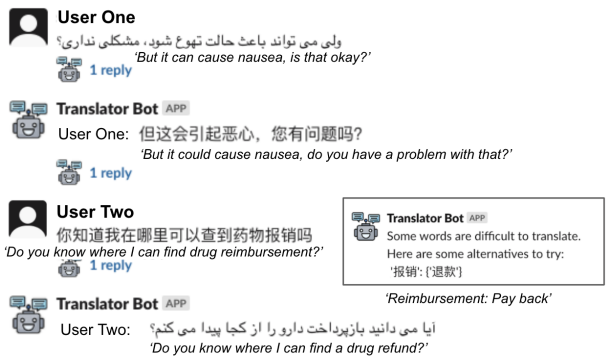

Figure 1: TranslatorBot mediates interlingual dialog using machine translation. The system provides extra support for users, for example, by suggesting simpler input text.

what kinds of information or interactions would best support user needs (Doshi-Velez and Kim, 2017; Miller, 2019). For instance, users may be more invested in knowing when they can rely on an AI system and when it may be making a mistake, than they are in understanding its inner workings.

Most consumer-facing translation systems do not provide support for safe and reliable use of machine translation. Despite some evidence of corpus-level human parity, individual translations still vary substantially in quality (Toral et al., 2018). Further, it is especially difficult for users to assess the quality of a translation because they often do not understand the source language, target language, or both. Mistranslations are frustrating for users (Hara and Iqbal, 2015), contribute to social and eco-

nomic isolation (Liebling et al., 2020), and have even prompted human rights violations, such as the wrongful arrest of a man whose social media post was mistranslated from "good morning" in Arabic to "attack them" in Hebrew (Berger, 2017). Nevertheless, people and organizations continue to rely on machine translation, even in high-stakes contexts like immigration (Torbati, 2019), and content moderation (Stecklow, 2018). To minimize the potential for harm from mistranslations, it is important that users are able to understand and account for the limitations of these systems.

In this paper, we explore what information and interactions can assist users of machine translation systems. Prior work has developed models to estimate the quality of a translation (Blatz et al., 2004; Specia et al., 2018). However, "translation quality" is a complex, nuanced, and context-dependent concept (Specia et al., 2013). It remains unclear when, how, and what kind of quality measures might be intelligible and actionable to users in different situations. We utilize HCI methods to understand what users need to know about translations in order to meet their goals. Our findings can inform the design of systems that support more effective use of machine translation.

First, we summarize key findings from a pilot survey of users' practices for evaluating machine translations. Informed by the survey findings, we propose three directions for the design of machine translation systems: 1) help users craft good inputs; 2) help users understand translations; and 3) adapt to context and feedback. To begin exploring these directions for design, we utilize a chatbot prototype that supports interlingual conversation mediated by machine translation (Figure 1).

2 Related Work

Increasingly, people use machine translation (MT) systems, such as Google Translate, to communicate across languages, from interacting with social media posts (Lim and Fussell, 2017) to negotiating employment or medical care (Liebling et al., 2020). Here, we focus on MT systems used by consumers, rather than systems designed for professional translators. These systems promise ease of access and improving quality (King, 2019), but it remains unclear how effectively these systems meet users' needs, particularly in transactional and conversational situations (Liebling et al., 2020).

Studies have shown that machine translation can improve communication in multilingual groups or teams (Wang et al., 2013; Lim and Yang, 2008; Calefato et al., 2016), but significant challenges remain. Poor quality translations can lead to frustration and conversational breakdowns (Yamashita and Ishida, 2006; Hara and Iqbal, 2015), and are detrimental to social and economic life for immigrants and people living and working in places where they do not know the dominant language (Liebling et al., 2020).

Researchers have found that people adapt to the limitations of translation models as they use a system, for example, by repeating, rephrasing, or supplementing information (Ogura et al., 2005; Hara and Iqbal, 2015). In some cases, users may treat a machine translation as a preliminary or gist translation, supplementing it with professional translations when needed (Nurminen and Papula, 2018). However, these strategies require users to assess the translation for errors and respond in a way that improves the translation output (Bowker and Ciro, 2019; King, 2019), both of which are challenging with affordances of existing systems.

One way to help users identify communication breakdowns and initiate repairs is to design systems that provide additional guidance and information. Researchers have proposed showing users alternative translations at the word or sentence level (Gao et al., 2015; Coppers et al., 2018), automatically providing back-translation (Shigenobu, 2007), displaying the estimated sentiment of a translation (Lim et al., 2018), or highlighting keywords (Gao et al., 2013). Experiments show that these approaches can improve message clarity, comprehension, and sense-making (Gao et al., 2013, 2015; Lim et al., 2018, 2019). Other research systems have attempted to detect cross-cultural translation issues using machine learning (Pituxcoosuvarn et al., 2020). Chatbots have also leveraged conversational content to translate (Hecht et al., 2012) or aid language learning (Cai et al., 2015). We build on this work and ask how we can guide users to identify when to rely on automated translations and when and how to make changes.

Researchers in NLP have developed methods to quantitatively evaluate MT models by comparing machine translations to reference translations (e.g. BLEU (Papineni et al., 2001)) or by Quality Estimation methods (QE). QE uses noisy parallel corpora annotated with quality information to estimate quality scores or classifications based on

linguistic and, more recently, model features (Blatz et al., 2004; Specia et al., 2020; Callison-Burch et al., 2012; Specia et al., 2018). Because QE does not require reference translations, these techniques could provide run-time quality estimates to end users. One challenge in QE is that evaluating translations is extremely complex and task-dependent (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010). Even a highly accurate confidence score or quality classification will never capture all the nuances of translation quality (King et al., 2003; Specia et al., 2013). One approach is to develop more complex QE models that combine multiple aspects of translation quality (Kane et al., 2020; Yuan and Sharoff, 2020). In this work we offer an alternative, human-centered approach: we begin by understanding users' needs and goals, and then explore what kind of quality information might be intelligible and useful to them.

3 Study: User Needs and Practices

When machine translation tools do not provide information to help a user evaluate translations, users must develop their own strategies for identifying breakdowns and initiating repairs. We conducted a survey to further understand this process.

We collected 267 responses to an English online survey, distributed through the authors' social media accounts in August 2020. The survey asked participants about how they use machine translation, how they evaluate translations, and how they respond to problems with translations (see supplementary material). We analyze 119 (44.5%) of these responses from people who needed translation at least multiple times per month and who used MT to meet those needs at least half of the time. Of those respondents, the median age was 29 years. 50% of respondents were men, 30% women, and 5.8% non-binary. The respondents collectively had some proficiency in 49 languages and used machine translation to translate between 44 different language pairs, 11 of which did not contain English. Most of the participants (114) used Google Translate, and 37 used MT on social media, e.g. translations provided by Twitter and Facebook.[1]

Limitations. Our survey was conducted in English and participants were recruited through convenience sampling online. For these reasons, our participants had high English proficiency, and most participants worked in research or the technology sector. Our results provide insight into the MT usage of these participants, and these results are strengthened by the wide range of languages that participants knew. Future work is needed to understand the needs and practices of other users, such as those with less familiarity with technology and those who are not fluent in high resource languages.

3.1 Results: Strategies for Evaluating Translation Quality

Consistent with prior work (Hara and Iqbal, 2015; Liebling et al., 2020), we found that poor quality translations are a problem for users of machine translation: 93% of respondents experienced poor quality translations in one or more contexts. Problems were especially pronounced in online contexts such as social media. Participants identified a number of known weaknesses of MT, including poor performance on informal or idiomatic language, domain-specific terms, longer passages of text, and text in low-resource languages or dialects (Nekoto et al., 2020; Luong et al., 2015; Bowker and Ciro, 2019).

The participants had various strategies for identifying problems with machine translations and recovering from errors. Their strategies highlight limitations of current MT systems for supporting informed and reliable use of translations. We outline three types of strategies participants reported, which inform our directions for MT system design.

Rely on (target) language proficiency. Identifying and responding to poor quality translations is extremely difficult without some proficiency in the target language. A commonly reported problem is inadequate translations that do not convey the meaning of the source text. This is not easy to identify if a user cannot understand both the source text and the translation. Participants also reported more nuanced problems such as formality inconsistencies and incorrect or unnatural sounding grammar.

Once users have identified a translation issue, strategies for repair are constrained by language proficiency. For instance, users can rely on their own knowledge of the target language to infer the meaning of an imperfect translation, or use their source language knowledge to modify the input text to improve the translation output. Asking a proficient speaker for help was a particularly effective and preferred strategy, but people using an MT system often do not simultaneously have access to

[1] Detailed demographics provided in Appendix B.

a proficient speaker in both languages.

Consult external resources. Some participants devised strategies to evaluate translations without relying on language proficiency. For example, participants used backtranslation or compared the output of multiple MT tools to estimate the reliability of a translation. Others referred to external sources of information, such as search engines, dictionaries, and encyclopedias. These strategies can be effective (Miyabe and Yoshino, 2009; Gao et al., 2015), but are time-intensive, and most require resources not provided by existing MT tools.

Use context-dependent quality assessment. In some contexts, users may choose not to evaluate translations. For example, one participant verified translations with a proficient speaker when at work, but not when on social media.

The key takeaway from the pilot survey is that users need more support to effectively utilize machine translation, including methods for preventing miscommunications, especially if they do not have proficiency in the languages they are translating between.

4 Directions for MT System Design

Based on the survey findings and pilot tests of TranslatorBot, our prototype chatbot to support MT-mediated conversation, we propose three directions for MT system design to help users more reliably meet their translation needs.

TranslatorBot is a command-based chatbot for interlingual communication (Figure 1).[2] The goal of the system is not only to translate between people writing in different languages, but to provide tools to help users avoid, identify, and recover from translation errors. The interface integrates translation support directly into dialog and the conversational nature offers flexibility to tailor interactions to the users' needs (Tsai et al., 2021). We have designed TranslatorBot to probe three directions for human-centered MT design:

4.1 Help users craft good inputs

Both quantitative (Lehmann et al., 2012) and qualitative (Bowker and Ciro, 2019) evidence suggests that MT models perform best on simple, succinct, unambiguous text ("controlled natural language"). However, this may not be clear to users who are not familiar with how these models work.

To improve translation quality, MT tools should identify messages that are difficult to translate and provide strategies to adjust the text. Similar techniques are common in search engines to help users craft better search queries (Morgan, 2010). However, crafting good candidates for machine translation is a much more complicated problem. In the MT setting, suggestions should be specific to the limitations of MT models, for example, simplifications or alternatives for idiomatic language.

TranslatorBot warns users when their input text is not a good candidate for machine translation. For example, TranslatorBot will warn the user if their message is very long, or if they used words that are unstable under backtranslation through high resource languages (Mehta et al., 2020). TranslatorBot also conducts a Google search on all input and uses the search engine's query suggestion (*"Did you mean . . . ?"*) feature to identify spelling errors or unusual language.

Open questions in NLP: How can we automatically identify when a text is a poor candidate for translation? How can we generate suggested improvements or alternatives? Some existing work has proposed predicting source words that lead to translation errors (Specia et al., 2018), but more work is needed to both identify when to intervene and to provide helpful, actionable, and effective suggestions to users.

Open questions in HCI: How can we teach users about the strengths and limitations of MT systems, in order to help them generate more "translatable" text? This work could build on existing literature in Information Retrieval that studies how to help users use search engines more effectively (Fernández-Luna et al., 2009). More broadly, HCI researchers have studied how to help users identify communication breakdowns and initiate repair (Ashktorab et al., 2019). How can we design natural interactions and interfaces to help users revise their original inputs when the current translation is causing confusion?

4.2 Help users understand translations

After providing users with a translation, MT systems should help identify errors and initiate repairs without assuming proficiency in the target language. Systems could leverage quality estimation (QE)

[2] See Appendix D for additional screenshots of the interface.

models to provide quantitative indicators of translation quality. In addition, systems can support and augment users' existing sensemaking strategies, for instance, by making it easy to view the backtranslation or a bilingual dictionary entry.

TranslatorBot provides users with more information about translations to help them make sense of what the other person is trying to communicate. Depending on the severity of the potential error, TranslatorBot might send this information automatically or users can request more information without leaving the conversation. The current prototype will warn the user if the literal translation of a unique word does not appear in the target translation, if the sentiments of the English translations of the source and the translation do not match up (Socher et al., 2013), or if a QE model (Kepler et al., 2019; Kim et al., 2017) predicts poor quality. Users can manually request a backtranslation, and we plan to integrate dictionary support in the future.

Open questions in NLP: How can quality information be better aligned with users' needs? For example, QE models have been developed at various levels of granularity (word-level, sentence-level, document-level (Specia et al., 2020)). How might these approaches be combined to provide useful, appropriate information based on the severity of errors and context of use?

Open questions in HCI: How can we design systems that help people make use of imperfect translation? In settings where people depend on translation for everyday tasks, a poor quality translation might be better than nothing (Liebling et al., 2020). What affordances might help users to make informed judgments about when to rely on a machine translation and when to seek alternatives?

4.3 Expand interactivity and adaptivity

Whether a translation meets a user's needs depends on the quality of the translation, as well as the context of use and the user's goals. MT interfaces, including signals of quality, should adapt to changes in context and to user preferences. One approach is to adjust the type and frequency of interventions based on contextual factors. For example, a user might prioritize accurate translation of domain-specific terms over fluency when using MT at a doctor's office. Future MT systems could leverage emerging techniques to give users more direct control over different aspects of translations, such as formality (Niu et al., 2017).

Conversational agents have great potential to support adaptive, interactive translation support. For example, TranslatorBot could be developed to ask users clarifying questions about their input. One case where this might be useful is to disambiguate input with multiple meanings. Users might also want to ask questions of TranslatorBot when a machine translation is unclear, or to make sure a translation is conveying their tone. Recent work in MT has studied whether providing more context directly to the translation model could improve translations (Specia et al., 2020). Integrating a translation system into dialog also offers opportunities for future work using conversation history as a source of context, both for machine translation and for integrated question answering tasks.

Open questions in NLP: How can we adapt Question Answering methods to support MT applications? How can machine translation models provide greater control to users over different aspects of a translation, such as tone? How can we utilize contextual information, such as a chat history, to improve the quality of machine translations?

Open questions in HCI: What aspects of a translation would users want control over? In what contexts might users have different needs and priorities for MT? Prior work in HCI has studied users' mental models of and preferred metaphors for conversational agents (e.g. (Khadpe et al., 2020; Ashktorab et al., 2019)). What roles or personas would users want a translation agent to play, e.g. interpreter, educator, confidence checker? How might this vary in different contexts? What kinds of questions would users want to ask of a translation system?

5 Conclusion

As the use of machine translation becomes more widespread, users need varied and reliable tools to assess translations in light of their specific communication goals. When considering what kind of information and interactions will be helpful to users, it is important to start from their varied needs, goals, and practices to ensure that systems provide intelligible and actionable support. As we have shown, interactive translations systems can support users to craft good inputs, make sense of resulting translations, and support them by adapting to context and user feedback.

References

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–12, Glasgow, Scotland Uk. ACM Press.

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.

Yotam Berger. 2017. Israel arrests palestinian because facebook translated 'good morning' to 'attack them'. *Haaretz*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland.

Lynne Bowker and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research*. Emerald Publishing, Bingley, UK.

Carrie J. Cai, Philip J. Guo, James R. Glass, and Robert C. Miller. 2015. *Wait-Learning: Leveraging Wait Time for Second Language Education*, page 3701–3710. Association for Computing Machinery, New York, NY, USA.

Fabio Calefato, Filippo Lanubile, Tayana Conte, and Rafael Prikladnicki. 2016. Assessing the impact of real-time machine translation on multilingual meetings in global software projects. *Empirical Software Engineering*, 21(3):1002–1034.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An Intelligible Translation Environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, Montreal QC, Canada. Association for Computing Machinery.

Michael Denkowski and Alon Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, CO, USA. AMTA.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

J. M. Fernández-Luna, J. Huete, A. MacFarlane, and E. Efthimiadis. 2009. Teaching and learning in information retrieval. *Information Retrieval*, 12:201–226.

Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 449–458, Paris, France. Association for Computing Machinery.

Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, pages 852–863, Vancouver, BC, Canada. ACM Press.

Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Kotaro Hara and Shamsi T. Iqbal. 2015. Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3473–3482, Seoul, Republic of Korea. Association for Computing Machinery.

Brent Hecht, Jaime Teevan, Meredith R. Morris, and Daniel J. Liebling. 2012. Searchbuddies: Bringing search engines into the conversation. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 6(1).

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral Based Interchangeability Assessor for Text Generation. *arXiv:2004.14667 [cs]*. ArXiv: 2004.14667.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW).

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Katherine M. King. 2019. Can Google Translate be taught to translate literature? A case for humanists to collaborate in the future of machine translation. *Translation Review*, 105(1):76–92.

Margaret King, Andrei Popescu-Belis, and Eduard Hovy. 2003. FEMTI: Creating and using a framework for MT evaluation. In *Proceedings of the Machine Translation Summit IX*, pages 224–231, New Orleans, LA, USA.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA. Association for Computing Machinery.

Sabine Lehmann, Ben Gottesman, Robert Grabowski, Mayo Kudo, Siu Kei Pepe Lo, Melanie Siegel, and Frederik Fouvry. 2012. Applying CNL authoring support to improve machine translation of forum data. In *Controlled Natural Language*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery.

Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet needs and opportunities for mobile translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13, New York, NY, USA. Association for Computing Machinery.

Hajin Lim, Dan Cosley, and Susan R. Fussell. 2018. Beyond Translation: Design and Evaluation of an Emotional and Contextual Knowledge Interface for Foreign Language Social Media Posts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12, Montreal QC, Canada. Association for Computing Machinery.

Hajin Lim, Dan Cosley, and Susan R. Fussell. 2019. How emotional and contextual annotations involve in sensemaking processes of foreign language social media posts. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Hajin Lim and Susan Fussell. 2017. Understanding How People Attend to and Engage with Foreign Language Posts in Multilingual Newsfeeds. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1).

John Lim and Yin Ping Yang. 2008. Exploring computer-based multilingual negotiation support for English – Chinese dyads: can we negotiate in our native languages? *Behaviour & Information Technology*, 27(2):139–151.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. 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 11–19, Beijing, China. Association for Computational Linguistics.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. 2020. Simplify-Then-Translate: Automatic Preprocessing for Black-Box Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8488–8495. Number: 05.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Mai Miyabe and Takashi Yoshino. 2009. Accuracy evaluation of sentences translated to intermediate language in back translation. In *Proceedings of the 3rd International Universal Communication Symposium*, IUCS '09, page 30–35, New York, NY, USA. Association for Computing Machinery.

Brian Stephen Morgan. 2010. Techniques for providing suggestions for creating a search query. US Patent 7,676,460.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages.

In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Mary Nurminen and Niko Papula. 2018. Gist MT Users: A Snapshot of the Use and Users of One Online MT Tool. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 199–208, Alacant, Spain.

Kentaro Ogura, Yoshihiko Hayashi, Saeko Nomura, and Toru Ishida. 2005. Language-Dependency in User's Adaptation for MT Systems in MT-mediated Communication. *Journal of Natural Language Processing*, 12(3):183–201.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press, Cambridge, MA.

Mondheera Pituxcoosuvarn, Yohei Murakami, Donghui Lin, and Toru Ishida. 2020. Effect of cultural misunderstanding warning in MT-mediated communication. In *Collaboration Technologies and Social Computing*, pages 112–127, Cham. Springer International Publishing.

Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. *Explanations as Mechanisms for Supporting Algorithmic Transparency*, page 1–13. Association for Computing Machinery, New York, NY, USA.

Tomohiro Shigenobu. 2007. Evaluation and Usability of Back Translation for Intercultural Communication. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, and Nuray Aykin, editors, *Usability and Internationalization. Global and Local User Interfaces*, volume 4560, pages 259–265. Springer, Berlin, Heidelberg.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. In Graeme Hirst, editor, *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Steve Stecklow. 2018. Why Facebook is losing the war on hate speech in Myanmar. *Reuters*.

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, Volume 1: Research Papers*, pages 113–123, Belgium, Brussels. Association for Computational Linguistics.

Yeganeh Torbati. 2019. Google says google translate can't replace human translators. immigration officials have used it to vet refugees. *Pro Publica*.

Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, page 1–13, New York, NY, USA. Association for Computing Machinery.

Hao-Chuan Wang, Susan Fussell, and Dan Cosley. 2013. *Machine Translation vs. Common Language: Effects on Idea Exchange in Cross-Lingual Groups*, page 935–944. Association for Computing Machinery, New York, NY, USA.

Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th anniversary conference on*

Computer supported cooperative work - CSCW '06, page 515, Banff, Alberta, Canada. ACM Press.

Yu Yuan and Serge Sharoff. 2020. Sentence level human translation quality estimation with attention-based neural networks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1858–1865, Marseille, France. European Language Resources Association.

A Survey Questions

1. How often do you need translation between two languages? (Almost never/About once a year/About once a month/Multiple times per month/Multiple times per week/Daily)

2. Think of a recent time when you needed to translate something. What was the situation?

3. In that situation when you needed translation, what did you do?

4. When you need translation, how often do you use an automatic translation tool (e.g. Google Translate)? (Never/Sometimes/About half the time/Most of the time/Always)

5. When you choose not to use an automatic translation tool (e.g. Google Translate), why? (I don't have access to any automatic translation technologies/I don't know how to use this technology/I don't trust the automatic translations/I don't need to use automatic translation technology (e.g. you have a friend or family member with you to translate)/Not applicable/Add your own)

6. Which automatic translation tools do you use? (Google Translate/iTranslate/Bing Microsoft Translator/Translations on Social Media/Translate Now/Add your own)

7. What kind of language do you translate? (Professional or formal/Casual/Intimate/Child-directed/Elder-directed/Domain-specific (e.g. medical or legal terms))

8. What are the most common settings where you use one of the translation tools? (For work/On social media/When traveling/To translate a website/During medical appointments/To learn a new language/Add your own)

9. When you use translation technology, what length is the text that you most often need to translate? (Single words/Single phrases/Sentences/Paragraphs)

10. How do you figure out whether a translation is correct or not? (Translating it back to a language that I know/Asking a native speaker/Breaking the translation down into single words or phrases that I can understand somewhat better/Looking up words in a bilingual dictionary/Checking whether it makes sense in context/I don't try to assess this/Add your own)

11. In which settings have you experienced poor quality translations when using an automatic translation tool? (For work/On social media/When traveling/To translate a website/During medical appointments/To learn a new language/Add your own)

12. Which of the following problems have you encountered? (The translation has a different meaning than the original text/The translation does not make sense/The translation has a different tone from the original text/The translation is too formal or too informal)

13. If you would like to add a problem you have encountered or expand on the problems you have encountered, please do so here.

14. What languages do you speak? For each language, please pick your proficiency level.[3] (Beginner/Conversant/Fluent/Native[4])

15. What language pairs do you usually translate between? (e.g. English-Mandarin, Spanish-Farsi)

16. What is your current occupation?

17. What is your gender?

18. What is your age?

[3]This question was presented as a grid with up to five rows where participants entered free response text for each language, and columns corresponding to proficiency levels.

[4]In consideration of accentism (https://accentism.org/) we will not use the term "native speaker" or "native" as a language proficiency category in future work. Future iterations of this survey will ask participants "How well do you read/write/speak your primary language?" with options: Not well at all/Somewhat well/Moderately well/Well/Very well.

Language	n
English	108
Spanish	53
French	44
German	30
Mandarin	15
Hindi	14
Japanese	14
Italian	10
Portuguese	8
Farsi	7
Arabic	6
Dutch	5
Russian	5
Swedish	5
Telugu	5
Turkish	5

Table 1: Languages that 5 or more participants have some proficiency in (Open responses to Q14: What languages do you speak?)

B Detailed Survey Participant Information

This section provides additional information about the 119 participants who were included in the analysis.

B.1 Language proficiency

Participants reported some level of fluency in 49 languages. Table 1 shows the languages that 5 or more participants listed. In addition, between 1-4 participants listed the following languages: Albanian, Amharic, ASL, Bangla, Cantonese, Croatian, Danish, Filipino, Finnish, Greek, Hausa, Hebrew, Indonesian, Irish, Kannada, Korean, Latin, Malayalam, Marathi, Marwadi, Norwegian, Pashto, Philippine Hokkien, Punjabi, Swahili, Tagalog, Tamil, Tigrinya, Urdu, Vietnamese, Wolof, Yoruba.

B.2 Language pairs

Table 2 shows response counts of language pairs reported by participants in response to Q15: "What language pairs do you usually translate between?" In addition to those shown in the table, each of the following language pairs was reported by one participant: Arabic-Farsi, English-Albanian, English-Bengali, English-Cantonese, English-Danish, English-Finnish, English-Gujarati, English-Latin, English-Malay, English-Norwegian, English-Tagalog,

Language Pair	n
English-Spanish	41
English-French	26
English-Mandarin	18
English-German	16
English-Japanese	16
English-Farsi	7
English-Arabic	6
English-Hindi	4
English-Korean	4
English-Portuguese	4
English-Russian	4
English-Dutch	3
English-Hebrew	3
English-Swedish	3
English-Amharic	2
English-Croatian	2
English-Greek	2
English-Italian	2
English-Swahili	2
English-Vietnamese	2
English-Yoruba	2
German-Portuguese	2
Japanese-Korean	2

Table 2: Language pairs that participants translate between (responses to Q15).

	Min	0.25	Median	0.75	Max
Age	18.0	25.0	29.0	36.5	74.0

Table 3: Distribution of participant age.

Occupation	n
Student	31
Researcher	22
Software Engineer	17
Professor or Lecturer	9
Consultant	5
Data Scientist or Data Engineer	5
Designer	5
Engineer	2
Other Technologist	2
Writer	2

Table 4: Survey participants' reported occupations. Open-ended responses were roughly grouped based on similar descriptions. Occupations reported by only 1 participant are excluded.

English-Telugu, English-Urdu, German-Spanish, Greek-Italian, Japanese-Mandarin, Mandarin-Hokkien, Portuguese-Spanish, Spanish-Arabic, Spanish-Basque, Spanish-Mapudungun.

B.3 Age

The age distribution is shown in Table 3. 10% of participants did not report their age.

B.4 Gender

Participants were asked "What is your gender?" with an open-ended response. 61 participants (50%) listed Male, M, or (Cis)male. 36 participants (30%) listed Female, F, or Woman. 7 participants (5.8%) listed Non-binary. Genders listed by only one participant are not reported for privacy reasons. 12 (10%) did not list a gender.

B.5 Occupation

Table 4 summarizes the participants' occupations. Occupations only reported by one participant are excluded for privacy reasons. 10% of participants did not list an occupation. Participants working in the technology industry and academia are over-represented, presumably because they were recruited through the authors' existing social networks. This is a limitation we are currently working to address in ongoing stages of this work.

C Translation Tools

Participants were able to select multiple translation tools that they use, as well as add their own open response. 113 participants selected "Google Translate" and one entered "Google chrome website translator (powered by google translate)." 34 selected "Translations on Social Media" and another 3 entered in the free response that they use translation tools on Twitter (2) and Facebook (1). 8 participants used DeepL[5]. 5 participants used Bing Microsoft Translator. Three participants used Apple products (e.g. translation app and Siri). 12 other services were listed by one participant each in the free response.

D TranslatorBot Prototype

Figures 2 and 3 demonstrate the current functionality of the bot prototype. The dialog is taken from a real conversation during a toy task where two of the authors role played a pharmacist and a customer and worked together to find the right medication for the customer, given their symptoms.

[5] https://www.deepl.com/en/translator

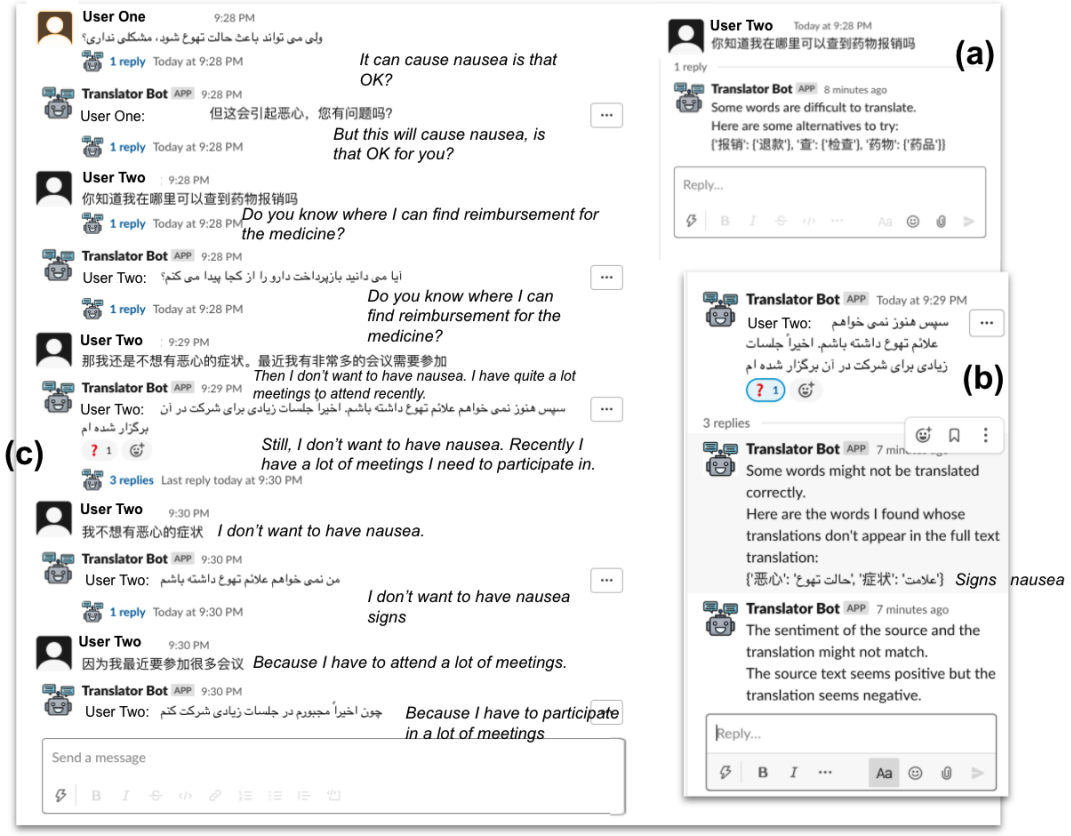

Figure 2: Two people use the TranslatorBot prototype to converse across languages. User One writes a message in Farsi, which is translated into Mandarin for User Two to read. When User Two responds in Mandarin, those messages are translated to Farsi so that User One can read them. TranslatorBot provides additional information in a comment thread: (a) if it detects that the input may be difficult to translate; or (b) if it detects possible problems with a translation. (c) User One indicated with an emoji reaction that the translation into Farsi was difficult to understand. This prompted User Two to break down their original message into shorter and simpler messages to resolve the communication breakdown.

Figure 3: TranslatorBot uses Google search query suggestions to help users catch typos before they are propagated through translation. In this example, User One has a minor typo in Farsi that dramatically changes the meaning of their message. The conversational context is a pharmacy setting, so a Farsi speaker may have been able to use contextual knowledge to identify the typo. However, if the incorrect text is translated, the person reading the translation may be unable to identify the source of the miscommunication. TranslatorBot shows the sender a private suggestion before translating the message, which the sender can accept or reject using emoji response buttons.